UNIVERSIDAD AUTÓNOMA DE MADRID

FACULTAD DE FILOSOFÍA Y LETRAS

Departamento de Lingüística General, Lenguas Modernas, Lógica y Filosofía de la Ciencia, Teoría de la Literatura y Literatura Comparada

MODALITY IN SPOKEN SPANISH AND JAPANESE: A CORPUS-BASED STUDY AND AUTOMATIC ANNOTATION

Carlos Herrero Zorita

Tesis doctoral dirigida por el Dr. Antonio Moreno Sandoval

2017

This dissertation was completed thanks to an FPI scholarship from the Universidad Autónoma de Madrid, and was carried out in the LLI-UAM Laboratory of Computational Linguistics, Department of Linguistics, under the supervision of Dr. Antonio Moreno Sandoval

To Carmen, and my parents.

De modalibus non gustabit asinus.

— Saying among medieval logicians

If life wasn't funny it would just be true, and that is unacceptable.

— Carrie Fisher

'Despair, or folly?' said Gandalf. 'It is not despair, for despair is only for those who see the end beyond all doubt. We do not.'

— J.R.R. Tolkien, The Lord of the Rings

Abstract

The main aim of this thesis is to automatically find and classify elements that signal modality in Spanish and Japanese sentences, taking into account both theoretical and empirical information. In order to join different disciplines such as typology, logic, corpus and computational linguistics, the aim is to answer three main questions: (1) What is the best definition and classification of modality for a cross-linguistic computational work? (2) How is modality used in spoken Spanish and Japanese, and how are modal markers modified in discourse? (3) How can this information be formalised into a program that can annotate modals automatically in new texts?

Modality is seen from the logic perspective as a semantic feature that adds necessity or possibility meanings to the predicate, as it is proven to be the best approximation for this type of study. Modality is encoded in the sentence in both languages by a series of auxiliaries, adverbs, adjectives and grammatical moods. The corpora will tell us how these markers are affected by negation, ellipsis, syntactic separation and ambiguity, which need to be detected by the program for the sake of precision and recall.

The corpora also provide information about modality usage, and reveals that it is a feature correlated to the type of communication, probably in relation to social constraints. Monologues achieve similar results in both languages, but when interaction takes place, the difference is noticeable. In dialogues, there is a predominance of necessity values in Spanish, and nearly equal numbers of necessity and possibility in Japanese.

The final result of the thesis is a rule-based program that outputs an XML with modal markers annotated and classified equally in both languages. It will be used in the future in bigger and different types of texts in order to draw more precise conclusions from both languages. Also, the program will be made available to use freely through an online interface at `http://elvira.lllf.uam.es/modtag/mainmodtagger.html`, hosted on the Computational Linguistics Laboratory web page of the *Universidad Autónoma de Madrid*.

Resumen (Spanish)

El objetivo principal de esta tesis es la búsqueda y clasificación automática de elementos modales en oraciones en español y japonés, usando para ello información teórica y empírica. En un intento de crear un estudio multidisciplinar entre tipología, lógica y lingüística de corpus y computacional, pretendemos responder a tres preguntas fundamentales: (1) ¿Cuál es la mejor definición y clasificación de la modalidad para un trabajo contrastivo computacional? (2) ¿Cuál es la frecuencia de uso en el español y japonés oral, y cómo el discurso modifica los elementos modales? y (3) ¿Cómo podemos formalizar esta información en un programa que pueda anotar automáticamente los marcadores modales en textos nuevos?

Consideramos la modalidad según la perspectiva lógica como un aspecto semántico que añade significados de necesidad o posibilidad al núcleo verbal. Se representa en ambos idiomas a través de una serie de auxiliares, adverbios, adjetivos y modos gramaticales. Los corpus nos dirán cómo estos elementos son afectados por la negación, la elipsis, la separación sintáctica y la ambigüedad, información que posteriormente será convertida en reglas a la hora de diseñar el programa y así aumentar su precisión y cobertura.

Los corpus también nos dan información acerca del uso y frecuencia de la modalidad en situaciones reales. Los resultados muestran que es un elemento de la lengua íntegramente relacionado con el tipo de comunicación, probablemente unido a las restricciones sociales. Los monólogos presentan unos resultados parecidos en ambas lenguas, pero cuando entra en juego una interacción, la diferencia es notable. En diálogos, la necesidad es el valor predominante en español, mientras que los hablantes japoneses usan casi de igual manera valores de necesidad y posibilidad.

El resultado final de la tesis es un programa basado en reglas que produce un archivo XML con los marcadores modales anotados y clasificados de la misma manera para ambos idiomas. El programa se usará en estudios futuros con datos diferentes y más extensos con el objetivo de confirmar los resultados obtenidos. Asimismo, estará disponible de forma online para su uso libre en `http://elvira.lllf.uam.es/modtag/mainmodtagger.html`, albergada en la página web del Laboratorio de Lingüística Computacional de la Universidad Autónoma de Madrid.

Acknowledgements

I would like to give a big thanks, first and foremost, to the Computational Linguistics Laboratory of the Autonomous University of Madrid which has made possible this incredible journey and provided me an enviable environment with countless of hours of knowledge, happiness (and headaches) during the last five years. To its director and my supervisor, Antonio Moreno for guiding me all the way through and providing me with help and answers to each and every question I made, no matter when I made them and how ridiculous they could be, and providing me with journeys around half of the globe. To the rest of the members of the team, a big thank you for being an example to follow, Alicia González, for sitting beside me in front of the computer day after day teaching me everything I know about Python and programming; Leonardo Campillos for all his wisdom and harmony; Antonio Pastor, IT magician who enrolled me into Linux and never said no when help was needed; and of course, José María Guirao from the University of Granada, always ready to share his knowledge.

This incredible journey at the laboratory would not have ever been possible without Clara Molina, the person who introduced me to it, and ignited an everlasting interest in language as an outstanding professor during my final years at the English Studies Degree. A special smile goes to Antonia Rodríguez Gago, another dear professor from the Degree who opened me to the world of theatre and almost, almost, lead me into studying a literature doctorate. I will, nevertheless, always remember her when watching in awe any Shakesperian play.

The second big deserved thank you goes to all the Japanese professors that have helped me throughout the harsh and savage journey that is Japanese learning during,

but also after, my years as a student at the UAM: Chieko Kimura and Emi Takamori, always welcoming with a big smile any problem I had; and Kayoko Takagi, for allowing me to take my first steps into teaching. To professors Toshihiro Takagaki, Hiroto Ueda and Shigenobu Kawakami, who received me with open arms in my research stay at Tokyo University of Foreign Studies in 2016, three of the most joyful months of my life, and showed me how an outstanding researcher can also be an outstanding person. Also, to professor Kiyoko Kataoka from Kanagawa University for all her patience during the long hot Friday afternoons in her office suffering the endless questions I made; and finally, professors Antonio Ruiz Tinoco from Sophia University and Ryuko Taniguchi and Sano Hiroshi from Tokyo University of Foreign Studies, for all their invaluable feedback.

An equally warming welcome was given to me during my research stay at Lancaster University in 2017. Thank you to professors Paul Rayson, Andrew Hardie, Mahmoud El-Haj, Scott Piao and all the people at the CASS centre for the wonderful opportunity at this incredible institution, and for all the professionalism, knowledge, wisdom and coffees granted to me during these three months.

Thank you to all those people that are battling this dragon too but always have time to lend a hand or an ear anytime anywhere, especially Carla Parra for her unending wisdom, Ryo Tsutahara for his help with Japanese and his love for life, and Yuanyi Liu for her positivism and happiness at the lab.

I have also to thank of course all my family and especially my parents for their unquestionable and blind support, and their weekly warm sheltering Sundays at home surrounded by unlimited affection and home-made meals that charged me up week after week.

And, saving the best for last, thank you to Carmen, unbreakable support and sister in arms in this absolute madness that is the PhD, who is always there for me, and without whom I could have never finished it.

Contents

List of Figures

List of Tables

List of Abbreviations

acc Accusative Case

adj Adjective

adv Adverb

aux Auxiliary element

cltc Clitic

cond Conditional

conn Connective

cop Copula

dat Dative case

elli Elliptic

emph Emphatic

expl Expletive

gen Genitive case

ID Identifier

imp Imperative mood

inf Infinitive form

irr Irrealis form

loc Locative case

MOD Modal Marker

neg Negative element

pln Plain form

pol Polite form

pot Potential mood

prog Progressive

prtp Participle

pst Past tense

quot Quotative

REF Referential

sbjv Subjunctive mood

te Te form

Chapter 1

Introduction

1.1 Purposes and motivations

This thesis will be devoted to a contrastive computational study between Spanish and Japanese of one of the most abstract aspects of human language: modality. In recent years, the area of modality has become extremely vague, with many different open discussions and positions depending on the background, discipline, or sometimes simply point of view, of the researcher. Linguistics is no exception. The modality from discourse analysis research may be very different from a purely syntactic one, and even more from the one presented in an applied linguistics work, such as language teaching, cognitive or corpus linguistics. Also, evidently, the modality understood by a European linguist will not agree completely with the idea held by a Japanese linguist.

There is, nevertheless, important common ground in this topic, something that every researcher on the area should agree with, or else we could not be talking about the single entity of *modality*. Modality is a unique feature to human reasoning and exclusive of human language. It is believed to be a connection between reality, mind, and word, and is present in every language of the world. It is closely related to tense and aspect, which also have a connection between the world that surrounds us and language, leading researchers to merge the concepts together. However, whereas tense and aspect are concerned with the *moment* of a certain state of affairs, with modality the speaker signals his or her *interaction* with these events, moved by a belief and a desire for it to become *true* or not. The main question is how is this connection made by the human mind realised in language, which has been answered in so many ways that the word of 'modality' itself is starting to dilute and lose its concrete meaning.

The main reason for this disparity has to do with the fact that modality is in essence a philosophical and psychological question that has been widely discussed for centuries, but has received the interest of modern linguistics only in the last decades. In Western studies, the issue began in Greece with Dionysius and Aristotle, and underwent a major breakthrough in modern logic and linguistics with Kant's ideas of judgement and proposition. In Japan it can be traced back to 13th century

scholars and poets, with its modern breakthrough in the *chinjutsu*, or the way the mind structures the predicates. The most common and worldwide definition of the term nowadays is the series of words a speaker of a language uses to display their *attitude*. The problem is, what is an 'attitude' of a speaker? Is it a personal opinion? Is it any kind of speech act? Is it factuality, or the encoding of certainty? Should we include human emotions in 'attitude'? Is it anything we as humans add to an objective proposition? Is the way we connect sentences in a text an 'attitude'? Is tense an 'attitude'? What about the elements we add according to restrictions imposed by society? Are the elements that signal politeness or gender 'modality'? Or is all of this false and the 'attitude' resides only in the grammatical 'mood'? The answer to all these questions is *yes*, depending on the nature of the study at hand.

When a computational linguist approaches this barren territory of conciseness and substantial thoughts, aiming to find clarity not only in one but two languages (three if we count English) there are two options, either to turn back and look for a clearer aspect of language, or face the problem. The study presented in these pages is the result of the second choice. The main idea is to select one interpretation of modality and find universal patterns that could be formalised in a computer code; in other words, to create rules that would allow us to automatically find modal elements in Spanish and Japanese texts and tag them with the proper categories. The research will try to follow three assumptions:

- *Universality.* Modality is believed to be at the same level as tense and aspect, a human element of language, so it is going to be present in Spanish and Japanese.
- *Simplicity before reach.* The study is situated in the corpus and computational linguistics area, so a simple classification that can be applied to both languages is more important than covering every single possible aspect of modality, even though some elements may be left out of the equation. In linguistic terms, we will only focus, for the time being, on overtly *marked* or grammaticalised elements in both languages that can transmit the idea of modality.
- *Balance between description and theory.* The ultimate element of this study, an automatic annotation of modal elements, must rest in a balanced structure

of theoretical and empirical data. Observation of how modality is used in real language using corpora is necessary for the development of the program, but this observation must follow a series of formal insights.

These ideas have driven us to consider modality to be based on two fundamental pillars: *necessity* and *possibility*. That is to say, it is formed by elements that claim a state of affairs to be true *in all possible worlds* (necessary) or true in at least *one possible world* (possible). The main classification will then be restricted to a binary choice, in an attempt to avoid as much ambiguity and obscurity as possible. Each pillar can also be either *epistemic*, if the interaction with the state of affairs is moved by a belief of the speaker; or *deontic*, if made by a desire. However, as soon as we try to look for more specific labelling we will find ambiguity, as we will see further on, making this subclassification not entirely reliable. Modality, is therefore a semantic feature of language, represented in the sentence by morphosyntactic elements, mainly auxiliaries and grammatical mood but also reinforced by adverbs and adjectives, in a similar way to tense and aspect.

This leaves out elements that are considered by many authors to encode modality. However, we move not only in a cross-linguistic work, the first one to include, to our knowledge, a quantitative comparison on the usage of modality between Spanish and Japanese, but also with a computational perspective in mind. Clarity is preferred before a complete coverage of the area, at least for now. As typologists, Bybee et al. (1994: 176) put it, "it may be impossible to come up with a succinct characterization of the notional domain of modality". Instead, the domain is usually characterized by referring to a set of more specific notions, each of which is defined separately, and which may be taken to share certain features motivating their grouping together under the label modality, but that will differ in many other respects. As such, the notion of modality is best viewed as a "supercategory" (Nuyts 2005), which is "much more loosely structured (and in fact probably belongs at a higher level of abstraction) than categories such as time and (types of) aspect" (Nuyts, 2006). An initial, clear and solid definition and classification of modality is essential before trying to move on to a further comprehensive study.

The study has three main objectives, each presented in a different chapter:

1. To find an appropriate definition of modality that can be adjusted to Spanish and Japanese in a computational study.
2. To perform a quantitative study of the elements that encode this feature using data from two spoken corpora to show an empirical distribution of the different categories.
3. To develop a program that could automatically find and classify these elements in future texts.

These are not at all independent, but build themselves upon each other. The study will begin with a theoretical discussion that will serve as the foundation of the annotation and analysis from the corpora. The development of the program is made taking into account these theoretical implications and the conclusions drawn from the corpora. The principal idea that moves this study is to join both theoretical and empirical information to create a series of rules and patterns that can be found in new texts and used to extract information.

The two main questions to answer in this dissertation then are: (1) How is modality used in both languages? (2) How can an automatic annotation of this feature be developed?

The principal motivation for this research is the apparent lack of corpus and computational linguistic studies comparing Spanish and Japanese modality. Separately, each language has an extremely long tradition of discussing mood and modality, but few studies have compared modality in general in both. The most singular studies are Wasa's comprehensive discussion (2005) and Fukushima's work (2013a; 2013b), although it only has a theoretical approximation and the definition of modality differs from the one taken in this study, featuring a clear Japanese perspective. No recent works have been found based on corpora comparing Spanish and Japanese modality.

There are, on the other hand, modality studies made with a typological or cross-linguistic view in mind (Bybee, 1985; Dahl, 1985; Hawkins, 1986; Bybee et al., 1994; Palmer, 2001; Van der Auwera et al., 2009; Horie & Narrog, 2014; Horie, 2014), which aim to find similarities in language with different features using scientific,

quantitative methods. This study will try to replicate this approximation focused only on the two aforementioned languages.

It has been also inspired by other automatic modal taggers (Lana-Serrano et al., 2012; Pakray et al., 2012; Morante & Daelemans, 2012; Rosenberg et al., 2012; Baker et al., 2015), although they have focused on a single language and have a more extensive view of modality. Similar to this study can be UCREL's Semantic Analysis System (USAS) (Rayson et al., 2004) which includes multilingual modality marking using the same annotation, although not including Japanese (Piao et al., 2015). There are also available several annotated corpora with some modal notions such as the corpus with evidentiality marking (Saurí & Pustejovsky, 2009) and semantic roles (Palmer et al., 2005).

1.2 Steps and structure

The steps taken for the development of this study are quite similar to the overall objectives:

1. First, to settle the definition and classification of the issue of modality used (Chapter 2). This will be done in three steps:
   (a) A comparative revision of the most important breakthroughs in Europe and Japan (Section 2.1). We believe the most appropriate way to approach the topic is to observe how the concept of modality has been formed across time.
   (b) A definition and justification of the position taken regarding modality and its classification (Section 2.2). This intends to be a typological and computational study, so the systematisation must be simple enough to avoid as much ambiguity as possible. It must be able to apply to both languages, and to be formalised into rules for the tagger.
   (c) An exposition of what is considered to be a modal marker, i.e. those elements that encode the notion of necessity and possibility (Section 2.3).
2. Secondly, to describe the methodology followed (Chapter 3):
   (a) The Spanish and Japanese corpora used for the annotation and quantitative study (Section 3.2). Each modal marker found in them has been annotated using XML tags and attributes, assigning modal information based on the theoretical conclusions such as main classification of modality (necessity/possibility), subclassification (epistemic/deontic/ambiguous), grammatical class (auxiliary, adverb, adjective or grammatical mood), negation and probability value, as well as possible features such as ellipsis or separation.
   (b) Description of the XML tagset used for the annotation in the corpora and listing of all the possible markers for Spanish and Japanese, tags assigned, analysis, and comparison between both languages. (Sections 3.3 and 3.5).
   (c) Description of the tools: XML language for annotation, Python for pro-

gramming and POS taggers for extracting morphological information from the text (Sections 3.3 and 3.4).

3. A quantitative study of the modal markers found in the annotated corpora, and main conclusions from the usage of modality in Spanish and Japanese (Chapter 4):
   (a) Comparison of the frequency distribution of modality and its types used by the speakers depending on language (Section 4.2), linguistic factors (discourse and register, Section 4.3) and non-linguistic factors (age and sex of the speakers 4.4).
   (b) Analysis of the elements that modify the modal markers like negation, separation of the auxiliary or the main verb, ellipsis of one of the elements, and possible errors made by the speakers in Spanish (Section 4.6).
4. Lastly, the development of a program that could replicate the annotation automatically in new given texts. The output is an XML tagged document with modal markers annotated with the tagset established previously, based on the theoretical insights. The program will also take into account the conclusions drawn from the corpus study, especially the modification of markers by a negative element or a separation in a sentence (Chapter 5).

1.3 Introduction to Spanish and Japanese

Before beginning with the study, a couple of notes should be made regarding Spanish and Japanese. The following pages from the main body of the study take into account some basic grammar knowledge of both languages that may take the unprepared reader by surprise. This section will try to describe in the most concise way the principal similarities and differences between both languages for the reader to properly understand the text.

1.3.1 Writing

Standard Spanish uses a single alphabet for its writing, the standard Latin system, with a series of additional characters such as the stressed vowels with the diacritic acute (´), the umlaut (¨) over the letter 'u' and the 'ñ' (*eñe*) letter.

Japanese, on the other hand, uses three different writing systems: *kanji*, *hiragana* and *katakana*. *Kanji* is a series of around 50,000 logographic characters adopted from Chinese characters that represent a morpheme or a phrase, mainly content elements. *Hiragana* and *katakana* are two phonetic-based syllabic writing systems, originated from the simplification of *kanji*, each character formed by one or two consonants and a vowel. The former is used primarily for inflection suffixes, function words and native words instead of *kanji*. The latter, *katakana*, is used for loanwords, foreign names onomatopoeia or emphasis.

Japanese characters and *kanji* can be transcribed into Latin characters (*romaji*). There are several romaji systems, but for this study I will be using the modified Hepburn system, which is characterised by the usage of a macron over long vowels and an apostrophe for the separation of the single 'n' syllable (ん) if followed by a vowel. Table 1 shows an example of the different writing systems in Japanese and their *romaji*:

Table 1: Writing systems in Japanese

English	Kanji	Hiragana	Katakana	Romaji
Grammar	文法	ぶんぽう		bunpō
Computer			コンピューター	kompyūtā

1.3.2 Grammar

In terms of grammar, there are three main differences between between Spanish and Japanese that are relevant for this study: word order, case marking and the inflection system. Spanish is a SVO language, sentences are normally structured as subject-verb-object. Japanese is SOV, the subject is normally in the head position and the sentence is ended with the verb –see examples 1 and 2, taken from Iwasaki (2013, p. 12). Japanese is a right-headed language, where the root or head is normally situated at the end, from word formation to language structure. However, the order is much more relaxed in Japanese, where permutation, ellipsis and pro-drop is very frequent in written and especially spoken language.

(1) *Un perro est-á com-iendo una manzana*
a dog be-PRES eat-PROG an apple

'A dog is eating an apple'[1]

(2) 犬 が りんご を 食べ-て-い-る
inu ga ringo wo tabe-te-ir-u
dog NOM apple ACC eat-TE-be-PLN

'A dog is eating an apple'

[1]The interlinear glosses in this study have been created following the Leipzig Glossing rules (*http : //www.eva.mpg.de/lingua/resources/glossing − rules.php*), albeit with some simplifications: the Rules establish a separation of all possible morphemes. Here, however, for the sake of clarity and ease of reading, we have focused exclusively on verb morphology, specifically tense and mood, the main focus of the study. For a complete inventory of the acronyms used, please consult the List of Abbreviations, pages XXV-XVI.

The second difference is case marking. Spanish, like English, has largely lost its case marking. It is indicated by phrase positioning or some pronouns, but does not have a specific affix or clitic to mark case. Japanese, on the other hand, has an overt case marking through postpositional clitics or particles. In the previous example, Example 2, the subject phrase is signalled by the nominative particle *ga* (が) and the direct object (accusative) with *wo* (を).

The third and final difference has to do with inflection. Japanese is a highly agglutinative language, not only in word composition, but also inflection. Spanish inflectional suffixes add tense, aspect, mood, person, number and voice to the verb. Japanese verbs, although lacking person and number inflection, do receive an extremely high amount of different grammaticalised suffixes and auxiliaries, from modal information to politeness, *gender-exclusive* forms, opinions, emotions, and so on. For example, as we will see, the Spanish auxiliary verb *tener* ('have', 'must', 'need') followed by a main verb in the infinitive and joined by the conjunction *que* form a modal periphrastic construction involving necessity and obligation (Example 3). The roughly Japanese equivalent is a series of grammaticalised auxiliaries (*nakerebanaranai*, なければならない) attached to the stem of a verb much like any tense suffix (Example 4).

(3) *Teng-o* *que* *qued-ar=me* *en* *casa*
have-PRES.MODAUX to.CONN stay-INF=CLTC at home

'I have to stay at home'

(4) 私 は 家 に い-なければならない
watashi *wa* *ie* *ni* *i-nakerebanaranai*
I NOM home LOC stay.IRR-have to.MODAUX

'I have to stay at home'

1.3.3 Register variation

Sociolinguistics focuses on the study of linguistic forms as the expressions of social relationships. This research is not located in this field. However, since part of modality has to do with interactions between human beings, a few remarks should be made on the topic.

In every communicative act, the *face* of the participants, or "public self-image every member of society wants to claim for themselves" (Brown et al., 1987, p. 61) is in play. As we will see below, modality is considered to express necessity and possibility values to the proposition. One of the ways of doing this is by imposing, through language, a state of affairs upon the receiver of the message, using, for example, a demand or a permission. These kinds of actions may *threaten* the face of the addressee by shaping their freedom of action (p. 65).

Any rational speaker will try to avoid or moderate these face threatening acts (FTAs) by using less direct linguistic forms. The level of moderation and face threatening depends on the relationship between two participants, and is regulated by cultural and social norms, restrictions and standards. Therefore, this will with no doubt influence the type of modal marker used if it involves the addressee. More specifically, there are two ways a modal marker can change in order to mitigate the FTA:

1. Through a different modal marker (lexical level).
2. Modifying the inflection of the marker (morphological level).

In the first case, in Spanish, for example, an imperative is more direct or face-threatening than a modal such as *deber* ('must', 'have to'). The same happens in Japanese, where a recommendation should be much more appropriate instead of a direct order (especially if the addressee is not close to the speaker) such as, 方がいい (*hōgaii*, 'should').

In other situations, a different inflection can be used to moderate the communicative act. In Spanish the conditional tense is used to metaphorically move the

context away and lessen an affirmation or demand (RAE, 2009, p. 473). The modal *deber* ('must') can be used in its conditional form *debería* ('should') for a more mitigated effect. In a similar way, the Japanese marker なきゃだめ (*nakyanadame*, 'must') sounds softer if ending *dame* is replaced with *naranai*.

The Japanese language, however, adds an additional complex of courtesy endings not found in Spanish. It is a language that changes substantially more than Spanish and English in terms of dialect, gender, age and, of course, courtesy. At risk of overgeneralising facts, since social norms depend on a large list of variables, traditionally Japan has been a society built on strong obligations upon the individual for the greater good of society. Individuals exist to ensure the well being, survival and self-sufficiency of the group, whether it is a family, a company, a city, or a whole country. Language serves as an undeniable medium for these obligations, resulting in the complex honorific system of *keigo*. It is a series of auxiliaries and lexical items used according to strict social rules that depend on the social status, age or closeness of the participants that is still used today. A complete discussion of *keigo* falls out of the scope of this study, but in everyday conversations we will commonly find basic polite variations in the inflection that must be taken into account. The most typical example is the suffix ます (*masu*), which can be added to improve the level of politeness, such as in the auxiliary できる (*dekiru*, 'can', informal) and できます (*dekimasu*, 'can', formal). As we will see further on, this feature will broadly extend the number of rules required for the automatic annotation.

Chapter 2

Modality

2.1 Definition of modality

When talking about modality, we face three main and dependent problems: its definition, its classification and its encoding via grammatical features of the sentence, also called *modality markers*. Each of them will be discussed in this chapter. Although the objective of this work is not focused on a detailed theoretical discussion regarding modality the multiple visions on the matter forces us to take a position among them and explain it before moving to the main study. This section will focus on the definition and classification, and Section 2.3 will describe the modal markers.

The most problematic issue concerning modality is the lack of a standard and unified position regarding its definition, classification and marking in language. As one of the major specialists of the area, Palmer (2001, p. 2) states, "in all typological studies there is considerable variation in the ways in which languages deal with grammatical categories, and there is probably more variation with modality than with other categories". The consequence is a tradition that spans several centuries of discussions and divergence of ideas on a topic nobody seems to be clear what it is. Since the 20th century and especially in the 90s, both Western and Japanese Linguistics have received countless of books and journals about the topic, scrutinizing the concept of modality to a point it has been diluted in dozens of study currents.

The clearest example in Spanish of this problem can be found nowadays in the reference grammar *Nueva Gramática de La Lengua Española*. First, the 'grammatical mood' (indicative, subjunctive or imperative) is defined as "one of the manifestations of modality" which "reveals the attitude of the speaker about the given information", but a few words below confesses that, however, the concept of 'attitude' is "imprecise" (RAE, 2009, p. 473)[1]. If we check its meaning of 'modality', a few chapters further on, it is defined as the *modus*, the part of the sentence that conveys the "attitude of the speaker" and stating that in the sentence 'Is it raining?' the *modus*, or modality, is the question. Hence, according to them, modality

[1] "El modo constituye una de las manifestaciones de la modalidad [...]. De acuerdo con la tradición, el modo revela la actitud del hablante ante la información suministrada, es decir, su punto de vista sobre el contenido de lo que se presenta o se describe. Se suele reconocer hoy, sin embargo, que aun siendo útil, el concepto de 'actitud' es impreciso" (Translation mine)

is also related to speech acts or illocutionary force of the sentence, whether it is an information conveyed or an *action* used in a social circumstance.(RAE, 2009, p. 794). Finally, we can find another issue of 'modality' in 'statement' or 'sentence' adverbs, those that indicate another *attitude* of the speaker, such as his or her certainty regarding a judgement. Therefore, modality is present at the same time in every sentence, the mood of the verb, the mind of the speaker, and some adverbs, but it appears to be considered to be too "imprecise".

This diluted and widespread understanding of modality has been caused by five main factors:

- Throughout history modality has been studied very closely with mood, joining them, in many occasions, as the latter being a representation of the former, leading to the conclusion that mood is a morphological codification of modality (Palmer, 2001; Moreno Cabrera, 2000). However, mood and modality have been studied not only from a grammatical point of view, but also philosophical, covering metaphysics, logic and philosophycal anthropology, leading many times to an entanglement of both disciplines.
- The concept of *mood* is per se problematic in linguistics. What exactly is represented by the mood of a verb? Is it the same the Spanish subjunctive mood, realised morphologically in the verb, and the English *subjunctive* created through the combination of different modal auxiliaries? What about the so-called *irrealis* inflection in Japanese, which cannot exist if not followed by a specific auxiliary? If they are different, are Spanish and Japanese *moods* and English *modality*?
- On the other hand, mood has not always been considered the same as modality. Both concepts have seldom been mixed between each other, merging *officially*, linguistically speaking, in the 20th century; but, there are still discussions on whether they should be separated or not.
- Although today's Western and Japanese linguistics have a similar identification of modality, the path that has lead to this point differs in a great way. We can trace back the concept of modality in European languages to figures like Aristotle, but Japanese linguistics do not have their roots in Ancient Greece.

Japanese modality is believed to have its origins in a psychological and almost poetical and metaphysical area of language, which eventually was labelled linguistically 'modality' (モダリティ, *modariti*) with the influence of studies Western scholars. Can we find a common ground between both cultures?

- The multiple studies about modality converged in Linguistics in the 20th century and reached more or less a common ground. However, in the most recent years, specially following Chomsky's theories, there has been another division, leading in some cases to contradictory theories covered by nearly all areas of Linguistics: Morphology, Syntax, Pragmatics and Semantics. The most common definition nowadays takes modality as the "opinion or attitude towards the proposition that the sentence expresses or the situation that the proposition describes" (Lyons, 1977, p. 452). This is, nevertheless, a very general and vague definition. On the one hand it is very convenient because it provides researchers a large margin of work, but on the other it lacks the specificity a work such as natural language processing needs. As Otaola Olano (1988, p. 435) explains, talking about modality in such general way may lead to misunderstandings. Modality acquires many interpretations depending on whether it comes from philosophy, logic, semantics, syntax, pragmatics, psychology or enunciation theory.

In other words, there is neither a clear origin of modality, nor a unified understanding or approach towards the term, both intra and cross-linguistically. This leaves the computational linguist in a rather complicated and dazzled, yet somewhat interesting, position. One should tread carefully when working in this area, explain the object of study as clear as possible, and try to not deviate from the path and get mixed up with other interpretations. This is the objective of this chapter, to explain how modality is understood for the purposes of this study. Since the aim here is to create a ruled-based system of modality for Spanish and Japanese which can automatically tag modal markers, I am interested in those features of modality that can be *seen* in the sentence. That is, we may be working with features that represent the *attitude* of the speaker, but we are only going to work with the *overt* or *marked* elements in the sentence that contain this information. Also, this appears to be the same stance taken by other similar typological, comparative and compu-

tational works. Nevertheless, an overview of the different positions on modality and some previous clarifications before explaining this decision seem necessary.

Due to its relation with mood, modality has been considered as a universal feature at the same level of tense and aspect, not only among languages but also among humans. It is a feature inherent to human knowledge and intelligence. Charles F. Hockett considered it along with 'temporality' the core of the 'displacement' property of human language, the one capable of referring to events not situated in the here and now (Von Fintel, 2006). Kant defined it as "one of the four classes of categories of human judgement", next to quantity, quality, and relation, which may be either a possibility, a necessity, or an existence (Van der Auwera & Zamorano Aquilar, 2015). Halliday (1970 [2009]) considered it the "social role" of the speaker, and "not a marginal language element, but one if its three primary functions, that concerned with the establishment of social relations and with the participation of the individual in all kinds of personal interaction". It is safe to assume that, as the marking of time, every natural language will contain a way to mark modality (Bybee et al., 1994; Palmer, 2001; Van der Auwera & Ammann, 2013).

If every language contains modality because it is inherent to human mind, a study of modality between Spanish and Japanese seems possible and justified. Nevertheless, these definitions are too general and are not sufficient for a computational work that looks for a concrete meaning and specific classification. In order to narrow down the issue and choose the appropriate stance a historical overview of the problem should be made, which will also serve as a summary of the most important works about modality. The review will be made as simple as possible, citing the most important works and milestones that have lead to the classificaiton used in this thesis. Also, since the issue here is to reach a definition that is optimal for both Spanish and Japanese, it is imperative to overview the development of modality studies from both Western and Japanese perspectives. Authors Grande Alija (2002); Narrog (2009a,b); Van der Auwera & Zamorano Aquilar (2015) provide a comprehensive review, from which my ideas have been taken.

2.1.1 First studies

The word 'modality' would not come until a fairly recent age in both Western and Japanese history. In the West, the issue has its roots in Ancient Greece, c.a. 2nd century B.C., with the concept of 'mood', and in this starting point we will already see a distinction in its definition.

We begin with two Greek Grammarians and two ideas. On the one hand, Dionysius Thrax described mood (*enklísis*, lit. 'disposition of the mind') in *Techné Grammatike* as a type of morphological attribute of the verb, and established five categories: defining, imperative, optative, subjunctive and infinitive. On the other, Protagoras defined mood as forms of discourse: wish, question, answer and command. In other words, Dionysius focuses on mood as a characteristic of the verb, and Protagoras as a characteristic of discourse.

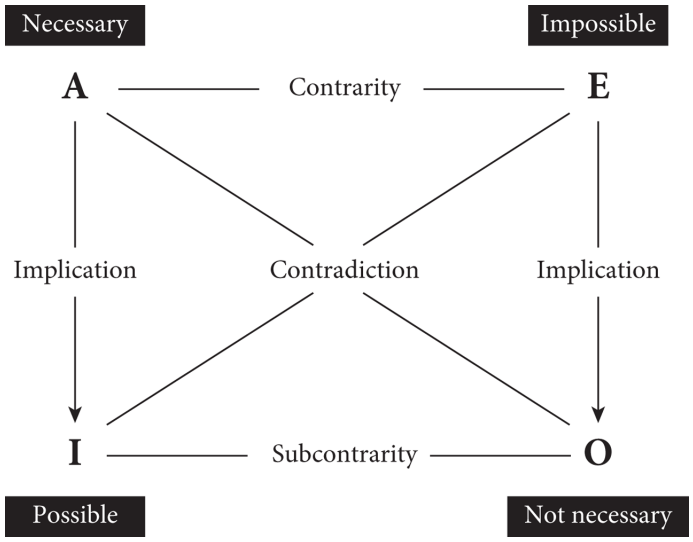
As well as grammarians, Greek philosophers were also interested on the topic. Aristotle's work *Analytica Priora* set the path of Modal Logic with his interest on the relations between subject and predicate (Patterson, 2002). He focused on syllogisms: drawing a conclusion via reasoning of two proposition or premises. Propositions must contain a subject and a predicate joined by the copula, and evaluate the affirmation or the denial of the predicate. Example 5 shows Aristotle's Barbara Syllogism (Rini, 2011):

(5) Barbara
A belongs to every B
B belongs to every C
A belongs to every C

He then introduces in these plain syllogisms the concepts of necessity, possibility, impossibility and contingency: the subject and predicate are necessarily or possibly conjoined with the copula. The aim of his work shifted to which pairs of propositions logically implied which conclusions including these new concepts. Years later, Apuleius of Madaura will be the first to represent the relationship of

these notions (Van der Auwera & Zamorano Aquilar, 2015), shown in Figure 1. An impossibility contradicts a possibility, just as necessity and non-necessity. But they are somewhat related: for example, a necessity of a proposition implies this proposition is possible. Possibility and necessity are contraries (both cannot be true) just as possibility and not-necessity are subcontraries (both cannot be false but can be true):

Figure 1: Square of Modality of Aristotle –taken from Van der Auwera & Zamorano Aquilar (2015).



These three definitions: mood as a morphological attribute of the verb by Dionysius, mood as forms of speech by Protagoras, and modal logic as the one concerned with necessities and possibilities of the proposition by Aristotle have remained for more than 2000 years, and form the basis of today's idea and discussion of modality.

The ideas from these philosophers were adapted to Latin, and studied in the Middle Ages, especially in the thirteenth and fourteenth century with the *modistae* and their Speculative Grammar. Precursors of the Universal Grammars, they understood language as a whole as a reflection of reality, conveyed through verbs, noun and adjectives, and introduced three modes or *modus* (Van der Auwera & Zamorano Aquilar, 2015):

- The first mode characterised the propositions as universal or particular, positive or negative.
- The second mode referred to the structure of the argumentation.

- The third mode, developed by Boethius' commentaries on Aristotle, divided the proposition into *dictum*, the propositional content (e.g. it is raining), and *modus*, which signals it as necessary, possible, impossible or contingent.

Surprisingly, roughly around the same time, there was also an increased interest in parts of speech in Japan, considered by many as the origins of today's understanding of modality. According to Senko (1993, p. 27), in the 13th century, scholar Fujiwara no Teika created a list of Japanese parts of speech: verb and adjective suffixes, particles, auxiliary verbs and conjunctions were labeled *te-ni-ha*, which later on will be considered as the encoders of modality; nominals, verbs and adjective stems, were named *shi*.

2.1.2 Birth of modality: mood among the inflection

In the West, the next big steps were taken in the Rennaisance with studies of the different European languages, including vernaculars. Scholars such as Antonio de Nebrija characterised mood as a category of the verb in Spanish, adapting the classification of Dionysius, and English and Portuguese grammars stated that mood was not only coded morphologically in the verb, but through auxiliaries now considered 'modal' like *can*, *could*, *might*, *dever* ('must'), *poder* (can), etc. Also, the term 'modality' made its apparition in French, *modalité*, as those aspects of a sentence that relate to its mood. (Van der Auwera & Zamorano Aquilar, 2015).

In the 17th century with the Universal Grammars and Philosophical Languages, theories of mood changed according to the language upon the grammar was being constructed, especially those that marked verbal mood through inflection and those that did not. James Harris' grammar adapted the Protagoras notion of mood as form of speech/text, including the interrogative. John Wilkins' grammar, one of the Universal Language grammarians, on the other hand, based on Latin and English, made another breakthrough: he took modal verbs as equivalents of verbal morphological mood that expressed notions of necessity and possibility (Van der Auwera & Zamorano Aquilar, 2015, p. 20,22).

Studies of different languages spread, and with this, in the next two centuries, the concept of modality and mood jumped between each other. The work that came from the Universal Grammars was focused on English, a language that lacks morphological mood, took Protagoras' discourse understanding of mood. Languages like Spanish have inflected mood, and its studies were closer to Dionysius' view (modality is seen in an attribute of the verb). However, these definitions did not remain the same and still kept changing. Grammarians of the 19th century understood modality as represented in Dionysius mood with Kant's reflections. He defined modality as one of the four classes of human judgement with three categories: problematical, assertorical and apodeictical:

> Problematical judgements are those in which the affirmation or negation is accepted as merely possible (ad libitum). In the assertorical, we regard the proposition as real (true). In the apodeictical, we look on it as necessary. Seen in Van der Auwera & Zamorano Aquilar (2015, p. 16)

Grammarians adapted his view, dividing modality into three moods, even in languages that did not represent it overtly as tense and aspect in the inflection of the verb: reality is expressed by the indicative, possibility by the conjunctive or subjunctive, and necessity by the imperative. In the West, modality and mood had reached a common position, though it would not remain as such for long.

At this time, as Maynard (1993) explains, in the Edo period of the 18th century Japan, the first of two seeds of what later become *modariti* for Japanese scholars had just begun to grow through the issue of *subjectivity* in language. Fujitani Nariakira updated the parts of speech classification by Fujiwara. Using poems for explaining his theories, he defended that a sentence was not only formed by parts of speech (*te-ni-ha* and *shi*), but also needed an 'echo' (*uchiai*), the series of personal feelings and emotions of the speaker that connected the words. A century later, Suzuki Akira defended that the *shi* words such as nominals, verbs and adjectives were contentful objects. The *te-ni-(o)-ha* words, like particles, were the 'strings' of the 'voices of the heart' that connected the *shi* contentful words and contained the attitude of the speaker.

The second seed arrived between the ending of the 19th century and beginning of the 20th. It was the concept of *chinjutsu* coined by Yamada, founder of modern Japanese Linguistics, and influenced by German and English psychologists of the time who were interested in defining sentence and the relations between subject and predicate. *Chinjutsu* stated that clauses are formed as the result of the psychological process responsible of the structuration and integration role of predicates (verbs and adjectives) in predicative clauses (declarative, interrogative and imperative) (Narrog, 2009b, p. 16). Though a different approach, this idea greatly resembles the initial definitions of mood in Western linguistics.

2.1.3 Consolidating the term *modality*

By the 20th century the discussion in the West of modality was opened again. Three main ideas remained intact:

1. Mood was a morphological feature of the verb (Dionysius)
2. Mood was a discourse feature (Protagoras)
3. Mood signalled necessity, possibility (Aristotle and Boethius)

Number (2) was now considered by many as the only way to express *modality*, and was represented by auxiliary elements that were different from the morphological mood. That is, verbal mood did not express modality, and they were two different things in a sentence (defended by Jespersen, Leech, Halliday and Coutes, among others). Other scholars such as Sappir and Zandvoot defended that both (1) and (2) were the same thing, modality was coded in morphological mood and auxiliaries (Van der Auwera & Zamorano Aquilar, 2015).

Number (3) was the stance taken especially by logicians with the birth of contemporary modal logic in the 60s. The most important discovery was the introduction of possible world semantics by Hintikka (1962) and Kripke (1963) that lead to the birth and separation of alethic logic, epistemic logic, deontic logic, etc. Many grammarians followed this understanding, as we will see below. Also, studies by Bally (1950) and Fillmore (1972) developed in France and America new ideas that

would greatly influence Russia and Japan's studies. Bally divided the sentence in two, separating the dictum (propositional content of the utterance) and the attitude of a 'modal subject' of a 'modal verb' that recreate the speaker's thought:

> The sentence comprises of two parts: the first is the correlative of the process of representation (e.g. the rain or a cure); we will call it the *dictum* as the logicians. The second contains the centrepiece, the one necessary for a sentence, corresponding to the process of the thinking subject. Modality is formed by the logical expression, the modal verb, and its subject, the modal subject (e.g. the one who believes, rejoices, wishes). Both form the modus, complementing the dictum –(Bally, 1950), seen in Zeman (2014, p. 469).[2]

Later, in a similar way, Fillmore stated the partition of the sentence, dividing it into proposition and modality:

(6) Sentence → Modality + Proposition. Seen in Narrog (2009b, p. 14).

In Japan, Yamada's *chinjutsu* was by the time developed even further by the scholar Tokieda, who took this concept and relabelled it as 'modality' (*modariti*), the element of the sentence that shows the speaker's *intention.* He identified *chinjutsu* as the *te-ni-o-ha* part of speech of Fujiwara, Fujitani and Suzuki, labeling them as *ji* (Maynard, 1993). Hence, for him the sentence was divided into the psycological, subjective voice of the speaker through the *ji* words, which signalled modality, and the *shi*, referential words that merely denoted concepts. Modality then was understood in Japan for several decades as a synonym of the speaker's subjectivity (Narrog, 2009b, p. 18).

[2]La phrase explicite comprend donc deux parties: l'une est le corrélatif du procès qui constitue la représentation (p. ex. la pluie, une guerison); nous l'appellerons, à l'exemple des logiciens, le dictum. L'autre contient la pièce maîtresse de la phrase, celle sans laquelle il n'y a pas de phrase, à savoir l'expression de la modalité, corrélative à l'opération du sujet pensant. La modalité a pour expression logique et analytique un verbe modal (p. ex. croire, se rejouir, souhaiter), et son sujet, le sujet modal; tous deux constituent le modus, complémentaire du dictum (Translation mine).

2.1.4 Modality today

Today, the discussion still lingers, mixing grammatical elements with philosophical and psychological issues. Are modality and mood the same thing? Or is mood a morpho-syntactic feature and modality a discourse-pragmatic one? Is it a philosophical or a linguistic problem? Is it a common feature of language, or is Spanish modality different from English and Japanese modalities? Hundreds of articles and books have been desperately looking for an answer and have not yet reached a common position. It will take time to reach a common understanding in an issue that from the beginning of linguistic studies has been understood differently. Any study, more even cross-linguistic ones, must take one of these ideas, the one that better adapts to the problem that wants to be addressed and follow it.

Today's works on modality fundamentally rest on two major definitions and studies of Lyons, Halliday, Lackoff and Palmer. The first described modality elements as those that "are used by the speaker to express, parenthetically, his opinion or attitude towards the proposition that the sentence expresses or the situation that the proposition describes" (Lyons, 1977, p. 451). Halliday (1970 [2009]) related modality to probability and certainty, and placed it outside the scope of tense. He understood modality as a subjective function of the speaker's language:

> Modality is a form of participation by the speaker in the speech event. Through modality, the speaker associates with the thesis an indication of its status and validity in his own judgement: he intrudes, and takes up a position. Modality this derives from [...] the 'interpersonal' function of language, language as expression of role. There are many other ways in which the speaker may take up a position, and modality is related to the general category that is often known a 'speaker's comment' [...], one among the syntactic complexes which together make up the interpersonal or 'social role' component in language. (Halliday, 1970 [2009], p. 176)

Lackoff considered modality to be present not only in grammatical elements but also in the performative meaning of predicates, and that should answer, for example, the difference between sentences 7a and 7b and 8a and 8b (Maynard, 1993, p. 34, 35).

(7) a. John says you must apologise. (Apology is required, speaker agrees with the proposition)

b. John says you have to apologise (The demand is reported, speaker does not need to agree).

(8) a. It's raining.

b. It's raining isn't it?

No modal marker is used, but the attitude of the speaker is present in both sentences. The speaker implies a personal guess, more explicitly marked in (b) with the expression 'isn't it?'

This was followed by Stubbs and Coates. Modality was related to the speaker's participatory attitude, and its commitment towards the proposition, from quoting a proposition (total detachment) to categorical asserting that it is the case (complete commitment). Stubbs (1986, p. 1) defined it as:

> Whenever speakers (or writers) say anything, they encode their point of view towards it: whether they think it is a reasonable thing to say, or might be found to be obvious, questionable, tentative, provisional, controversial, contradictory, irrelevant, impolite, or whatever. The expression of such speaker's attitudes is pervasive in all uses of language. All sentences encode such a point of view,... and the description of the markers of such points of view and their meanings should therefore be a central topic for linguistics –Taken from (Maynard, 1993, p. 35).

These considerations by Lyons, Halliday and Lackoff's have given scholars and linguists the opportunity to explore modality freely, leading to different approaches and discussions regarding modality. In the last decade modality has become deeply rooted in pragmatics joining its definition to Austin's *speech acts* theory (1975), discourse analysis, sentiment analysis, Conversation Analysis (Cepeda & Poblete, 2006) and textual dimension. Research has been particularly focused on the definition and discussion of modal markers, specially on their textual and rhetoric functions (Cornillie & Pietrandea, 2012).

Another issue that has been greatly discussed in the last decade is the one related to evidentiality and modality. Evidentidality is considered a 'linguistic cat-

egory whose primary meaning is source of information [...] To be considered as an evidential, a morpheme has to have "source of information" as its core meaning' (Aikhenvald, 2005). Since some of the evidentials express the possibility of an event by the opinion of the speaker, it is only natural that it will overlap with the previous definitions of modality. Discussions in the last years have therefore pondered the idea of whether or not evidential markers must be included in the same category as modal markers and viceversa (De Haan, 1999; Cornillie, 2007, 2009, 2010; Squartini, 2004; Hennemann, 2013).

My approach to modality is more related to typological work such as Palmer's (2001). Originally published in 1986, it is closer to the logicians tradition, reducing modality to a dual choice between necessity and possibility. This belongs to another trend related to modality that has taken part in linguistic and typological research, based on quantitative and comparative studies across languages. He also performed a cross-language study, in the same way as other authors such as Bybee, Perkins and Pagliuca (1994), Van der Auwera and Ammann (2013) and Plungian (1998). In this line of work modality is seen as those elements of the sentence that encode the necessity or the possibility of the proposition, heavily influenced by modal logic theory. It is marked in the sentence by grammaticalised morphemes such as auxiliaries and affixes that may include morphological moods.

Studies on modality in Japan have been also greatly influenced by these approaches of the second half of the 20th century. Modality as the present modern concept properly began in the 60s and 70s combining Yamada and Tokieda's ideas of *chinjutsu* and Bally and Fillmore's segmentation of the sentence. Modality was considered as those elements that are added to the proposition and carry the speaker's subjective stance (Narrog, 2009b). However, once again, scholars did not develop a unified position towards this matter, and several currents of thought took place and are still carried on today. On the one hand, Masuoka (1987) created a dual view, understanding modality as everything marked outside the proposition. He and Takubo (1992, p. 117) defined it as 'mood' (ムード, *mūdo*), stating that:

> In the situations when a sentence is used as a communicative tool, the speaker does not only express a specific state of affairs (SOA). He or she is also expressing at the same time a judgement or an attitude to the

> addressee and the state of affairs. The speaker can let the addressee know about a SOA he or she believes in (assertion); request some information from the addressee (question); demand things (imperative, prohibition, request); a necessary judgement made by the speaker induced by a SOA (obligation); to express the knowledge that cannot judge the truth (general remark); the judgement of a negation (negation); to explain the differences between two SOA (explanation) or the characteristics of two similar SOA (comparison). In this way, *mood* is all the grammatical forms that expresses the speaker's attitude or judgement towards the SOA and the addressee[3].

These forms are, in summary, all the different sentence modifiers added by the speaker, which are situated at the end of a sentence: the conjugated forms of the predicate, auxiliary verbs, ending particles, and other ending forms (Masuoka & Takubo, 1992, p. 117).

Nitta (1985), on the other hand, adapted Lyon's view that modality was the speaker's attitude that subjectively modified the proposition, but proposed a multilevel layering instead of a two side division like Masuoka. For both of them, modality was an indespensable semantic element of sentence formation, present in all sentences (Narrog, 2009b, p. 12).

Japanese modality studies are primarily based on the ideas of these two authors. However, in the last decade some have criticised this idea, although a minority. Precursors of this position are Onoe (1990) and Nomura (2003). The former, influenced by Langacker, views modality in verbal mood, that is, modality is inside the predicate, represented by elements in predicative form (auxiliaries and complex endings) that describe an irrealis state of affairs. The latter criticises modality as a subjective marking and defines it as the expression of the relationship between sentence concepts and reality (Narrog, 2009b, p. 30). Closer to our view, Harada (1999) and Johnson (2003) situate modality on the axis of necessity and possibility, but their influence among the Japanese tradition is quite dim (Narrog, 2009b, p. 32).

[3]話し手が、文をコミュニケーションの道具として使う場合、ある特定の事態の表現だけではなく、その事態や相手に対する話し手の様々な判断・態度が同時に表現される。それはある事態を自分の信念として相手に知らせるものであったり（確言）、相手に情報を求めたり（疑問）、聞き手に対する様々な要求であったり（命令、禁止、依頼）、あり事態が生じることの是非に関する話し手の判断であったり（当為）、真とは判断できない知識を述べたり（概言）、否定の判断であったり（否定）、ある事態で特徴づてかり（比況）、といったものである。事態や相手に対する話し手の判断・態度を表す文法形式を一括して「ムード」と呼ぶ (Translation mine)

Table 1 will summarise the most important milestones of the history of Modality. Of course, there have been hundreds of studies on the topic, with many important authors discussing it. This is but a summary of what we consider to be the most relevant discoveries that could clarify the concept of Modality for this study. For a comprehensive historical and comparative review, consult Grande Alija (2002); Narrog (2009a,b); Van der Auwera & Zamorano Aquilar (2015).

If we combine today's main approaches to modality in Western and Japanese linguistics, we can group them in three main trends (updated from Moriya & Horie (2009, p. 97) and Cornillie & Pietrandea (2012, p. 2109)).

a) Modality is everything modifying the proposition, including negation, tense, case particles, discourse markers, etc., present in every sentence: (Fillmore, 1972; Masuoka, 1991; Givón, 1995; Wasa, 2005; Nuyts, 2006; Imithani, 2009)
b) Modality is the expression of the attitude or subjectivity of the speaker, also his or her emotions and opinions:(Halliday, 1970 [2009]; Lyons, 1977; Nitta, 1991; Bybee et al., 1994; Palmer, 2001)
c) Modality relates language with reality: expression of necessity/possibility, factuality, realis/irrealis in either the morphological mood, the modal auxiliaries, or both: (Givón, 1995; Harada, 1999; Johnson, 1999; Palmer, 2001; Nomura, 2003; Narrog, 2009a)

The first two, a and b, are the most widespread currents in Japanese modality. In the West, the majority of studies move around currents b and c. Each way of understanding modality is of course perfectly valid, but working in this area inevitably means taking a position on the matter and picking one of the sides. The selection and marking will rest entirely on the nature and objective of the study at hand. It is interesting, however, how these lines of study appear to have an element in common: the presence of human mind realised in language and the relation between its production and our understanding, as humans, of reality. Nevertheless, this definition of modality still fails to be concrete enough, and it is necessary to select one of these currents and narrow it down for our study.

Approaches a and b, often linked together as many linguists consider the ele-

Table 2: Most important historical shifts in linguistic Modality studies in West and Japan

Century	West	Japan
4th-2nd BC	**Aristotle** - Modal Logic. Necessity and Possibility in syllogisms. **Dionysius** - Mood was the disposition of the mind in a morphological attribute of the verb. **Protagoras** - Mood as forms of discourse: wish, question, answer, command.	
13th	**Modistae** - Greek ideas from Latin + Theology. Language as a reflection of reality. Parts of speech and modes. **Boethius** - Dictum and modus.	**Fujiwara** - Japanese parts of speech. *Te-ni-ha* + *shi*.
14th-17th	**Universal Grammars** and Languages - Mood in "auxiliarised" verbs or verb inflection. Represents notions of modal logic.	
18th-19th	Grammarians take **Kant**'s view on modality: indicative, conjunctive and imperative moods represent reality, possibility, necessity.	**Fujitani** - Sentence formed by *Te-ni-ha* + *shi* and the *uchihai*, feelings and emotions of the speaker. **Suzuki** - The *shi* were conveyers of content, the *te-ni-(o)-ha* words contained the attitude of the speaker that connected the *shi*.
19th-20th	Scholars **joining Modality** and Mood - Sappir, Zandvoot. Scholars **separating Modality** (modal auxiliaries, speech acts) and Mood (inflection of verb) - Leech, Halliday, Palmer, Coutes. **Logicians** following Boethius' modus signalling possibility and necessity. **Bally and Fillmore** - Separation of sentence. Modality + Proposition = Sentence. **Lyons** - Modality expresses the attitude of the speaker. **Halliday** - Modality as the subjective function of the speaker's language. **Palmer** - Logicians tradition, modality is reduced to Necessity, Possibility.	**Yamada** - *Chinjutsu*. Psychological process responsible for the integration of the predicate in the predicative clauses. **Tokieda** - *Te-ni-(o)-ha* renamed as *ji*. The *ji* words contained the *chinjutsu*, signaled modality. The *ji* denoted concepts. **Masuoka** - Influenced by Tokieda, Fillmore and Bally: separation of sentence. **Nitta** - Influenced by Tokieda and Lyons: the speaker's attitudes subjectively modified the proposition.
21st	**Pragmatics** - Modality as speech acts, discourse markers, signals text coherence, sentiment analysis. **Typology** - Modality as a universal grammatical category marking possibility, necessity, factuality, certainty, evidentiality, etc.	

ments outside the proposition as the ones that mark the speaker's subjective stance (especially among Japanese scholars), are very non-effective computationally speaking. The problem with the subjective approach to modality (approach b), is that

the vagueness of its definition has lead to many interpretations. It may be useful for studies based on pragmatics or communicative intentions and discourse analysis, but if we want to consider it as a feature represented syntactically like tense or aspect, it is non-viable (Grande Alija, 2002). As Narrog (2009a, p. 4) puts it, "if taken seriously, it leads to a disproportionate expansion of the category and potentially even the absorption of most other grammatical categories such as voice, aspect, tense or illocutionary force, which also tend to be strongly associated with the attitude of the speaker". Selecting this option would lead us to pages of discussion trying to narrow down the subject and probably leaving many elements aside. Or, to an eventual *overtagging* of modality, including functional elements such as discourse markers or Japanese particles, to annotating the illocutionary force of the elements. Even further, if we follow the traditional approach of Japanese scholars considering modality as a fundamental and obligatory element of the sentence, we would be forced to tag *all* the sentences of a corpus, something that may not be possible in every domain or not true in other languages such as Spanish. Finally, the creation of hand written rules for an automatic tagging of the subjective voice of the speaker would be virtually impossible. The solution would be to manually tag those elements of the sentences that could encode subjectivity from the speaker, train a program, an test it in new data. This, however, would require very large amounts of data and the learning algorithm would be reduced to a series of probability calculations made by the machine, without allowing us to extract linguistic patterns.

Option c seems the proper option as it has already been used in previous typological studies. Nevertheless, is still a very wide area, it does not provide us with a proper understanding of the phenomena and needs some narrowing down.

Considering previous cross-linguistic studies on modality, this study will be influenced by the ones carried out by authors such as the already mentioned Bybee (1994), Palmer (2001) and Van der Auwera, Plungian and Ammann (2013; 1998). All of them have something in common: they refer to grammatical markers like mood inflection suffixes and auxiliaries, and they are based on the logical tradition of considering modality as a dual paradigm of necessity and possibility. The advantages

of following the logic tradition is that it is not restricted to a language but to a way of reasoning of the human mind, which gives us a good opportunity to apply it to two very, apparently, different languages such as Spanish and Japanese. It is important to try to start from a universal perspective instead of a specific language, especially if one of them is our mother tongue, as it may bias our view. For example, taking English modal verbs as starting point for defining modality as the majority of linguistic studies did at the end of the 20th century may risk leaving modal elements not present in languages such as Japanese or Spanish because they do not exist in English and vice versa.

These linguists' understanding of modality has its roots in modern modal logic, which was born on the second half of the 20th century from the long logician tradition originating in Ancient Greece (see previous section) and follow a semantic/syntactic approach. It follows the tradition of alethic logic, or the one that understands that a proposition representing the state of affairs may be qualified as either necessary or possible. The next section will further describe this position.

2.2 Selection of the appropriate modality

As we have seen in Section 2.1, modality can present a challenge for the computational field. Not only because of its variety of definitions, but because the majority of them view it as an abstract element that may present a problem if we want to develop context-independent classification and rules. Between the approaches a, b, c, I have considered the last one, that is, modality as the expression of elements such as necessity and possibility, as the most appropriate for the task.

More specifically, modality is a mental process signalled by the speaker that considers a state of affairs to be true in every possible situation (necessity) or true in at least one possible situation (possibility). Due to its popularity in modality study, this will be followed by a subclassification between epistemic (if this consideration is made by a belief of the speaker), deontic (made by a desire) or ambiguous (both). Each level of classification will be described in the following sections.

2.2.1 First level: Necessity and possibility

As stated in Section 1.1, the final objective of this work is twofold:

- To perform a comparative quantitative study between Spanish and Japanese modality based on spoken corpora.
- To develop an automatic modality tagger for both languages based on rules extracted from the theoretical and empirical information.

Since we are working in a computational field, we are dealing with the two fundamental challenges any study in this area will encounter to achieve the highest level of precision and recall: ambiguity resolution and portability. As Abney (2011) explains, any natural language is filled with ambiguity. For example, automatically assigning a morphological category to the word *duck* in the following sentences, could prove to be very bothersome –From Abney (2011, p. 4):

(9) a. When he began flailing about, he made her duck.
b. When he invited her to dinner, he made her duck.

For a competent speaker of English with some linguistic knowledge, it is easy to tag the *duck* from sentence (a) as a verb and the *duck* from b) as a noun, but to formalise it for a computer is not so straightforward. There are two solutions for this problem: (1) the human creates the rules for the computer to understand it, a solution very popular in the 80s, or the most modern one, (2) make the computer create its own rules via machine learning.

Machine learning, although extremely powerful as it is, does not provide the linguist any linguistic explanation for the solution, only probability calculations. A solution for studying modality can be to manually annotate what we consider the attitude or subjectivity of the speaker and then allow the computer to automatically learn. This is a possible solution for trends a and b, as we saw in the previous section. Not only we will obtain a *useless* method linguistically speaking, as the problem of formalisation will remain unsolved for the humans; also the issue of subjectivity can be very difficult to define, and due to its vagueness, an annotator may understand it different from another.

The creation of rules allows a solution of the problem based on our linguistic knowledge. The formalisation is made by the human, and then processed by the computer. The downfall of this approach as it has been revealed in the last decades, is its inability to deal with situations impossible to standardise. Natural languages are ambiguous and complex and constantly changing, and many situations would require too many rules to process them. If working with this approach, the studied feature should as much contained and objective as possible.

The second challenge mentioned by Abney is portability, or "the difficulty of porting a system developed for one subject domain to a new domain" (2011, p. 4). As shown by Biber (Biber, 1991; Biber et al., 1999), language frequencies and variables change according to the type of text and discourse. Any kind of study has to take this into account, and computational ones are no exception. If the approach

is rule-based, the rules should be contained enough to adapt to different situations. If it is generated through machine learning, the training process must be repeated in different types of texts.

This study aims to tackle these problems with the creation of hand-made rules. The focus is ruled-based, as we want to formalise the coding of modality and create a series of the instructions based on observations in theoretical studies and corpora for a tagger to automatically annotate modal markers. Therefore, the understanding of what modality is, and the way it is coded in the sentence, must resolve these challenges as efficiently as possible. Modality encoding in the text may be ambiguous: one marker can denote several types of modality, as I will explain below; Also portability, because we are moving between two languages, registers (formal, informal) and discourse types (monologues, conversations and dialogues). Therefore, our understanding, definition, classification and marking of modality must fulfil the following requisites:

1. It must be accountable for the grammatical differences between Spanish and Japanese.
2. It must have a morphological and syntactic approach, moving away from pragmatics.
3. It must work independently from context.
4. It must classify modal markers avoiding as much ambiguity as possible, providing a sufficient amount of relevant information.
5. It must be compatible with other elements present in the discourse like negation or ellipsis.

The best way to approach this will be through modal logic, as it easily resolves ambiguity, portability and formalisation. Also, one of the most successful and widespread applications of modern logic, specially mathematical logic, has been the development of computer and computer programs. The best way to formalise modality into rules will be through the formal aspects of modality based on logic.

The most common definition of modal logic (as, once again, there is a lack of consensus on the matter), is the so-called alethic logic, the one that understands that

the truth value of a proposition representing the state of affairs may be qualified as either necessary or possible, expressed through modal markers such as adverbs (possibly, necessary, etc.) or auxiliaries (must, may, etc.). It rests on the philosophical studies pioneered by Aristotle and Boethius, and adapted by many typological linguists, previously explained. Considering the following sentences:

(10) a. It may rain tomorrow.
b. You must eat more vegetables.
c. I am possibly mistaken.

Sentences 10a., 10b. and 10c. can be represented respectively by the following formulae selecting between *possible* and *necessary*:

(11) a. The fact that tomorrow is raining is possibly true.
b. The fact that you eat vegetables is necessarily true.
c. The fact that I am mistaken is possibly true.

Since alethic modality is a propositional or sentential logic, as it studies the modification of propositions, in this case, using necessity and possibility, we can express the formulae with symbols. If a proposition p is necessary, it is represented as $\Box$p. If a proposition p is possible, it is represented as $\Diamond$p. This can be applied to any language, which makes it very attractive to cross-linguistic studies. Sentences 12 and 13 show and example in each language.

(12) 明日　　は　雨　かもしれない
ashita　*wa*　*ame*　*kamoshirenai*
tomorrow NOM rain may.MODADV

'It may rain tomorrow'

(13) *Probablemente*　*lluev-a*　*mañana*
Probably.MODADV rain-SBJV tomorrow

'It (will) probably rain tomorrow'

Sentences 12 and 13 can both be formulated as 'the fact/truth value that tomorrow is raining is possible', or simply '◇p', being p the proposition 'rain tomorrow'.

The notions of *necessity* and *possibility* may lead to misunderstandings. To better explain the concepts, we must address Kipler's 'possible worlds' (1963), the understanding that, at least abstractly, an infinite number of worlds, universes, or state of affairs is possible at any moment. In our case, we are evaluating the speaker's utterances; hence, the set of possible worlds (w) is established by him/her. The fact that it may rain tomorrow is established by the speaker, according to his/her own knowledge.

Logic assumes each sentence is either true or false (the Law of Excluded Middle), but not both true and false (the Law of Non-Contradiction) (Kaufmann et al., 2006). If the truth value of a proposition is necessary (□p), it is true in all possible worlds. If the truth value of a proposition is possible (◇p), it is true in at least one of the possible worlds. The set of possible worlds where the proposition is true has been called 'modal base' (R) (Kratzer, 1981). Taking a sample sentence (φ), and (V) as the evaluation function (0 for False, 1 for True), this can be formalised as the following (taken from Kaufmann et al. (2006, p. 80):

(14)

$$V_w(\Diamond_p\varphi) = \begin{cases} 1 \text{ if } V'_w(\varphi) \text{ for some } w' \in p_w \\ 0 \text{ otherwise} \end{cases}$$

$$V_w(\Box_p\varphi) = \begin{cases} 1 \text{ if } V'_w(\varphi) \text{ for all } w' \in p_w \\ 0 \text{ otherwise} \end{cases}$$

In addition to this, the issue of negation can also be easily processed. As we saw in Figure 1 of Section 2.1 by Apuleius, necessity and possibility are connected, and can be implied through one another, by negation. Adding negation to the operators change them to the opposite operator:

(15) a. $\Box$p $\iff$ $\neg\Diamond\neg$p

b. $\Diamond$p $\iff$ $\neg\Box\neg$p

That is, 'necessary p' is equivalent to 'not possible not p': 'It will rain tomorrow' if and only if 'it is not possible not to rain tomorrow'. Whereas 'possible p' is equivalent to 'not necessary not p': 'It may rain tomorrow' if and only if 'it is not necessary no to rain tomorrow'. In other words, the negation of a possibility becomes a necessity in the form of an impossibility, whereas the negation of a necessity becomes a possibility as a 'not necessity' implies the possibility or the event becoming true, or not. Another example can be seen in the following sentences taken from Palmer (2001, p. 91):

(16) a. Mary must come tomorrow. - Necessity

b. Mary may come tomorrow. - Possibility

c. Mary can't come tomorrow. - Not-possibility, i.e. necessity. (Mary not coming tomorrow is necessarily true)

d. Mary needn't come tomorrow. - Not-necessity, i.e. possibility. (Mary coming tomorrow is possibly true)

Therefore, the first level of the tree that forms our classification of modality will be divided into Necessity and Possibility. If the modal marker states that the sentence is true in one of the worlds perceived by the speaker, it will be tagged as Possibility. If the marker on the other hand states the sentences will be true in every possibility, it will be tagged as Necessary. The next subclassification will consist on Epistemic or Deontic modality, and will be explained in the following section.

2.2.2 Second level: Epistemic and deontic

One of the most popular classifications of modality today shared by many distinguished authors of the area, even with a typological view in mind, divides it into two categories that come from the subfields of modal logic with the same name: deontic modality and epistemic modality (Von Wright, 1951). Due to the great amount of relevant studies covering both types, it seemed reasonable to include them in the classification, as they are perfectly compatible with the necessity and possibility distinction. However, their tagging is secondary in terms of importance. Even though they provide a more specific meaning to the verb, they bring additional ambiguity for precisely the same reason, and the results must be treated with caution.

Epistemic[4] modality[5] expresses the degree of probability of the state of affairs according to the knowledge of the speaker. In other words, it indicates an estimation made by the speaker of the chances that the state of affairs expressed in the sentence applies in the world (Nuyts, 2006, p. 6). De Haan (2006, p. 29) defines it as the "degree of certainty the speaker has that what s/he is saying is true". Palmer (2001) describes it as "the speaker's attitude to the truth-value or factual status of the proposition". The following sentences show an example of epistemic necessity (17) and epistemic possibility (18).

[4]From Greek *episteme* 'knowledge' (Kaufmann et al., 2006, p. 103)
[5]Also called 'impersonal' or 'propositional' (Palmer (2001); RAE (2009, p. 573))

(17) Tom must be at home (The fact that Tom is at home is necessary)

(18) Tom may be at home (The fact that Tom is at home is possible)

Deontic[6] modality[7] refers to the "degree of moral desirability of the state of affairs expressed in the utterance, typically, but not necessarily, on behalf of the speaker" (Nuyts, 2006, p. 4), and deals with features such as obligation and permission. As with epistemic modality, we can include it inside the values of necessity and possibility, paraphrased with 'possible' and 'necessary' (Examples 19 and 20 –taken from Palmer (2001):

(19) Kate must come in now (It is necessary for Kate to come in now)

(20) Kate may come in now (It is possible for Kate to come in now)

The main problem of this classification is the famously known ambiguity or overlapping between both types. In other words, the same modal markers are used for epistemic and deontic values. Taking for example the following sentence 21:

(21) John may enter the room

This sentence can be understood either as a permission (the speaker allows John to enter, deontic reading) or a possibility perceived by the speaker (the speaker knows that it is possible that John is entering the room, epistemic reading). This overlapping can be found in many languages of the world, and is present also in Spanish and Japanese (Grande Alija, 2002; Van der Auwera & Ammann, 2013; Narrog, 2012; Akiba, 2014). However, as we will see further on, the amount of ambiguity between epistemic and deontic forms in Japanese is extremely low, barely perceivable, compared to Spanish, mainly due to two reasons: (1) Japanese contains a much higher array of modal markers, each with more specific meanings, and (2)

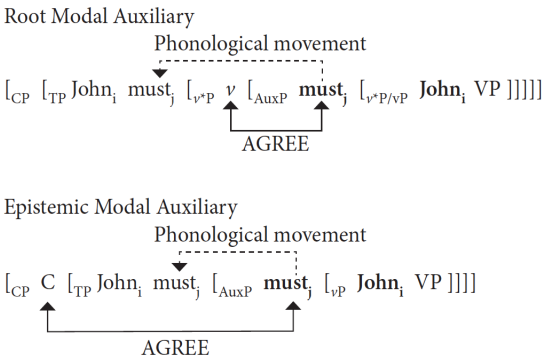
[6]From Greek *deon* 'obligation' (Kaufmann et al., 2006, p. 103)

[7]'Root modality' for some authors (Larrega, 2009), 'radical'. 'personal' or 'situational' for others (RAE (2009, p. 573); Van der Auwera & Ammann (2013))

certain modal expressions restrict the subject of the sentence. That is, the speaker maintains a position when expressing private emotions and affairs, but when he or she wants to express someone else's, they have to express them in a different way and hence use another marker (Nariyama, 2003, p. 89).

When using an ambiguous marker, the semantic choice between them depends entirely on the intention of the speaker, which may, or may not, signal it in the context of the sentence. Some studies (Nuyts, 2001; Clinque, 2006; Akiba, 2014, p. 22) have revealed that they have different syntactic interpretations, not semantic. Regarding its scope among the sentence, epistemic markers modify the whole sentence. Their scope is over the subject but they are outside the scope of tense. Deontic markers, on the other hand, have scope only over the predicate. The scope of subject and tense is above them. For these reasons, Palmer (2001) names them Propositional Modality and Event Modality, respectively and so do Van Valin and LaPolla (1997). To put it in other words, as Akiba (2014) states, from a Chomskian's syntactic perspective, a deontic modal is situated in the VP phase (hence the labeling '*root* modal'), whereas an epistemic one would be in the CP phase, outside the proposition, and outside the scope of tense. Figure 2 (p. 21) represents this claim.

Figure 2: Syntactic position of deontic (root) and epistemic modals



This high level of disambiguation can prove very difficult for the human without the proper context, even more if a spoken conversation is being analysed, as sometimes the participants will share common knowledge and it would not be necessary for the speaker to be explicit with the information shared. This difficulty becomes nearly impossible for the computer through formal rules. Utterance 22 is taken from the corpus as a real life example of ambiguity between epistemic and

deontic modality:

(22) *Tú no pod-rías trabaj-ar en el Gran Hermano ese tía*
you NEG can-COND.MODAUX work-INF at the Big Brother that mate
porque todo el día est-arías pendiente de la cámara
because every the day be-COND waiting for the camera

'You couldn't work at Big Brother mate because you would be looking after the camera all day' (i.e. 'it is impossible for you to work' (epistemic) or 'you are unable to work' (deontic)

In fact, since the ambiguity is syntactic, it means a modal marker can have two interpretations at the same time. As Stowell (2004) presents in his example (23), ambiguous modal markers can have two different meanings at the same time, depending on the syntactic analysis, which can also be observed in the previous example (22):

(23) *El ladrón pud-o entr-ar por la ventana*
the thief can-PST.MODAUX enter-INF through the window

'The thief was able to enter through the window' (deontic)

'It is possible that the thief entered through the window' (epistemic)

There are, nevertheless, some semantic restrictions between them. For example, a deontic marker is not compatible with impersonal sentences, or if the values of 'capacity', 'disposition' or 'intention' cannot be assumed by the subject (RAE, 2009). An example of these are sentences with atmospheric verbs, such as 24a and 24b:

(24) a. *Pued-e* *llov-er* *más*
can-PRES.MODAUX rain-INF harder

'It can rain harder' ('It has the ability to rain harder')

b. *雨 が 降-る-ことができる
ame ga fu-ru-kotogadekiru
rain NOM fall-pln-can.MODAUX

'It can rain more' ('It has the ability to rain')

In Spanish, the perfect infinitive can give us another clue: if the auxiliary is in its present form, the periphrasis will only be compatible with a perfect infinitive main verb with its epistemic reading, as in the sentence:

(25) *Pued-o* *hab-er=lo* *escri-to*
can-PRES.MODAUX have-INF=CLTC write-PRTP

*'I can have written it' (Only grammatical with the epistemic reading in Spanish)

Some authors explain that this ambiguity is related to the diachronic semantic changes in modality, a process that tends to shift from non-modal lexical elements to speaker-oriented, deontic modality and then to epistemic meanings. Van der Auwera & Plungian (1998) drafted a semantic map of modality representing this change which has recently been updated (Van der Auwera et al., 2009) (Figure 3). However, authors such as Narrog (2012) believe this is not sufficient for other

languages such as Japanese as there are markers that do not follow these directions, and the modelling of semantic changes should address wider notions as shown in Figure 4.

Figure 3: Van der Auwera and Plungian's modality's semantic map (updated) (Van der Auwera et al., 2009).

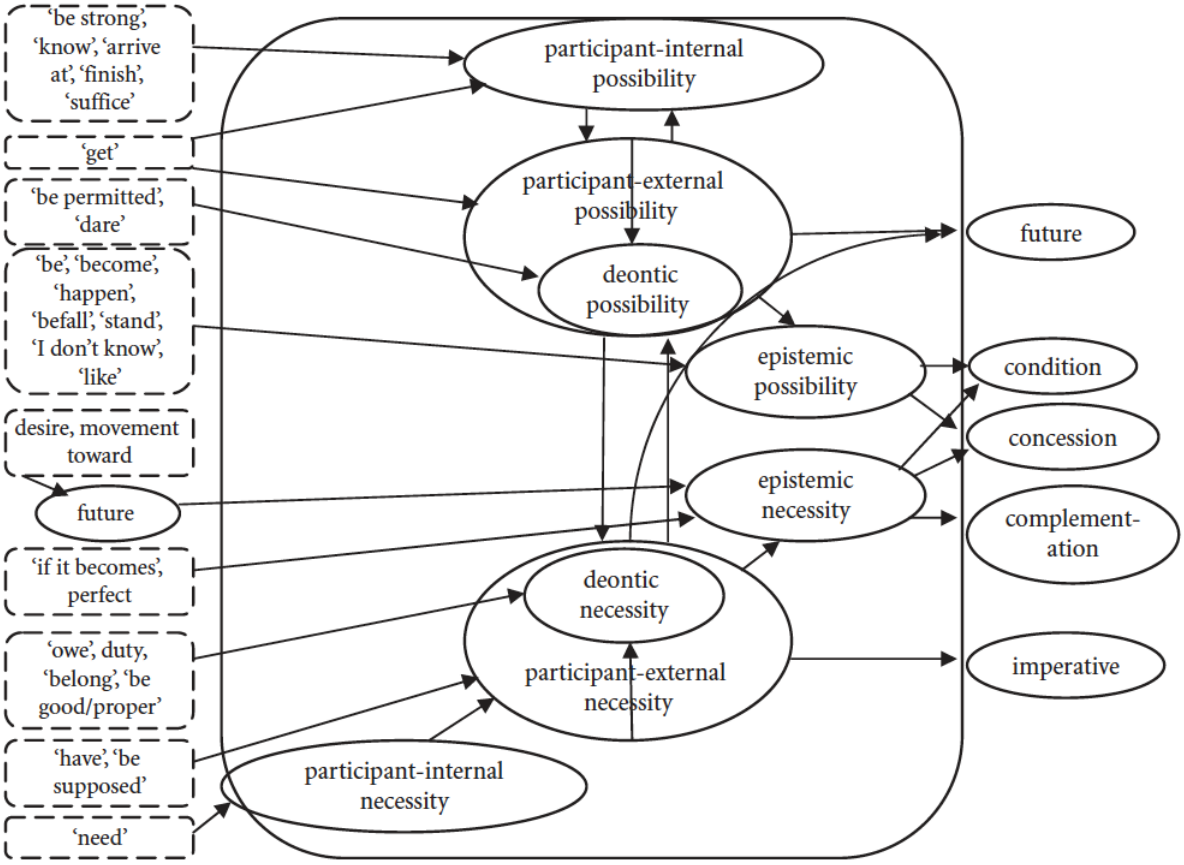
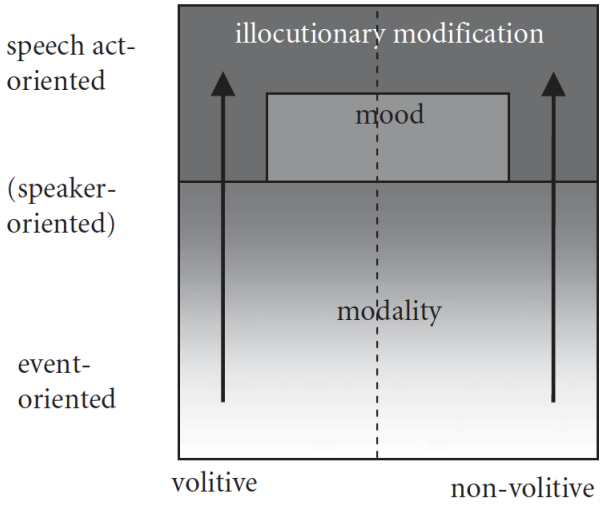


Figure 4: Narrog's modality map (Narrog, 2012).



Some of these selective features can be applied as rules for the automatic tagger to reduce some ambiguity. However, for a comparative and automatic study, the main distinction must be between necessity and possibility, as it does not present any ambiguity problems. A separation between epistemic and deontic will also be included in this study, but, those markers that can pose some overlapping will simply remain ambiguous. Not only the ambiguity is difficult to separate, even for a human, in many cases it is impossible to separate it, as both readings are possible, depending on the syntactic analysis chosen. Grande Alija (2002, p. 66) states that "relying on the Necessity/Possibility pair will allow us to establish a unity and a consistency. This way we can resolve the serious problem of vague openness and dispersion of modality"[8]. Kaufmann et al. (2006, p. 80) give us another example. Considering the following sentences:

(26) a. John is at the party
b. John may be at the party
c. John must be at the party

Both (26b.) and (26c.) "can be used as assertions about either the speaker's beliefs (given what I know...) or John's options and obligations (given the rules and John's age...), but the logical relations invoked in the two sentences are unambiguously possibility and necessity, respectively".

2.2.2.1 Other classifications

Although many wide-known authors follow the insights of logicians necessity and possibility, the majority focus on a classificaiton of Epistemic/Deontic. Others, especially those with a typological view in mind and not satisfied by these definitions, have developed further classifications that will not be used here but are worth mentioning.

[8] "Apoyarse en el binomio de Necesidad/Posibilidad ayuda a consolidad la unidad y trabazón. Se consigue así solventar el grave problema de la amplitud y dispersión de la modalidad". (Translation mine)

Bybee (1994), with a diachronic study in mind, proposes four types of modality: agent-oriented modality, which "reports the existence of internal and external conditions on an agent with respect to the completion of the action expressed in the main predicate", indicating obligation, necessity and ability (which includes root or deontic modality); speaker-oriented modality, which includes directives such as commands, demands, permissions, etc.; epistemic modality, signalling probability, possibility or certainty; and subordinating moods, complement clauses, concessives and purpose clauses that mark modality.

Van der Auwera and Plungian (1998) developed a semantic map and classification of modality based on Bybee's ideas, which has influenced this study. They too consider modality as a dual paradigm of necessity and possibility, and they develop several kinds of modality: epistemic modality; participant-internal modality (the possibility or necessity internal to a participant engaged in the state of affairs: the participant's ability for possibility, the participant's internal need for necessity) and participant-external ability (refers to circumstances that are external to the participant engaged in the state of affairs and make this state of affairs either possible or necessary), which subdivides into deontic or non-deontic modality. Figure 5 represents their classification:

Figure 5: Van der Auwera and Plungian's modality types (1998, p. 82)

<table>
<tr><td colspan="4">Possibility</td></tr>
<tr><td colspan="3">Non-epistemic possibility</td><td rowspan="2">Epistemic possibility
(Uncertainty)</td></tr>
<tr><td rowspan="2">Participant-internal possibility
(Dynamic possibility, Ability, Capacity)</td><td colspan="2">Participant-external possibility</td></tr>
<tr><td>(Non-deontic possibility)</td><td>Deontic possibility
(Permission)</td></tr>
<tr><td rowspan="2">Participant-internal necessity
(Need)</td><td>(Non-deontic necessity)</td><td>Deontic necessity
(Obligation)</td><td rowspan="3"></td></tr>
<tr><td colspan="2">Participant-external necessity</td></tr>
<tr><td colspan="3">Non-epistemic necessity</td><td>Epistemic necessity
(Probability)</td></tr>
<tr><td colspan="4">Necessity</td></tr>
</table>

Van der Auwera and Plugian's study also supports my idea of locating epistemic and deontic's values under necessity or possibility. For example, a permission or a capacity are types of possibility, in the same way as need or obligation are types of necessity. I will, however, limit the branches and levels and simplifying them into epistemic, deontic or ambiguous. A preliminary study was made on the Spanish cor-

pus, distinguishing between participant-internal and participant-external modality (Herrero, 2014). However, the overlapping became even more serious and rendered it unproductive.

Palmer (2001) develops even further the concepts of epistemic and deontic modality and establishes two more modalities. On the one hand, inside propositional modality and aside from epistemic modality, there is evidential modality. The difference between each other is that "epistemic modality speakers express their judgments about the factual status of the proposition, whereas with evidential modality they indicate the evidence they have for its factual status" (p. 8). On the other, event modality divides into deontic and dynamic modality. Whereas in deontic modality the conditioning factors are external to the individual, the factors that condition dynamic modality are internal (p. 9). Similar to participant-internal and external modality from Van der Auwera & Plungian (1998).

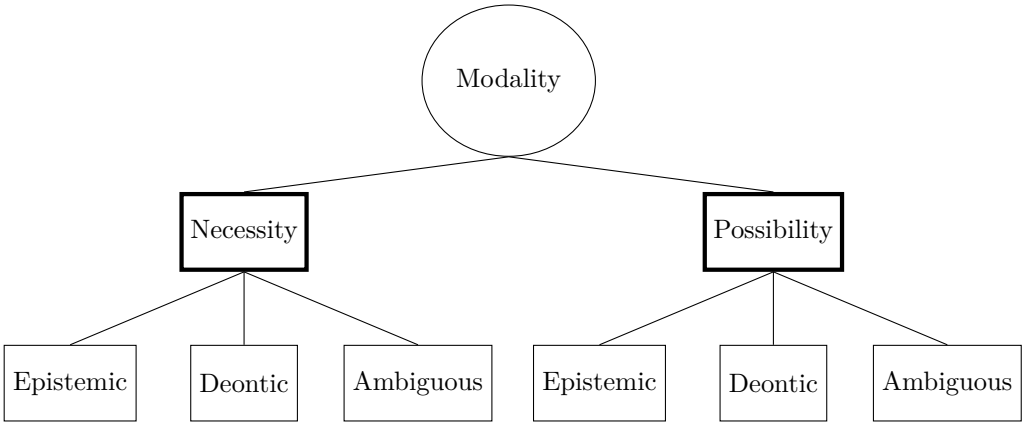
Looking at the different array of interpretations and classifications of modality, even only selecting the most influential studies, it is understandable how the topic has expanded so much nowadays. As Bybee (1994, p. 176) states at the beginning of the chapter on modality:

> Mood and modality are not so easily defined as tense and aspect. A definition often proposed is that modality is the grammaticization of speakers' (subjective) attitudes and opinions. Recent crosslinguistic works on mood and modality, such as Palmer's, however, show that modality notions range far beyond what is included in this definition. In fact, it may be impossible to come up with a succinct characterization of the notional domain of modality and the part of it that is expressed grammatically.

2.2.3 Putting everything together: the modality used in this study

Any of the presented definitions and classifications of modality is equally valid, one must simply choose the one that satisfies the purpose of the study. Here, the classification of modality is summarised in the following tree (Figure 6).

Figure 6: Classification tree for modality used in my study



Hence, what does a modal marker represent: the realisation in language of the necessity or possibility of a reality element or state of affairs (SOA) in the speaker's mind, either a non-controllable event (epistemic modality) or a controllable one via the hearer (deontic modality). Modality, then, is a reflection of how the speaker, a human, interacts through language using his mind with everything he or she perceives in the, so to speak, *real* world, the one that occurs out of his/her mind. Van der Auwera (1985, p. 27) draws an interesting diagram that represents this idea (Figure 3). We can divide the speaker's mind (M) in two: an interactive device (ID), which interacts with the out of mind (OOM) world and the storing device (SD), which does not interact with the OOM. The SD consists on beliefs and desires, the ID on consciousness and intentions.

Table 3: Four-dimensional model of the mind.

MIND			
SD		ID	
Beliefs	Desires	Consciousness	Intentions

A belief is a reflection of a SOA in the process of conceptualisation (a SOA generated in the Mind). It does so by creating an object of consciousness, generated with the interaction of the ID with the OOM, and transmitting its conceptualisation to the SD. A desire on the other hand is the cause of the reflection of a SOA conceptualisation. It this SOA is set in the future, the desire may result in an

intention. Van der Auwera (1985) defines each feature as the following:

1. A belief is a lasting SD-object resulting from an attempt of the ID to reflect a SOA in a conceptualization by creating a momentary object of consciousness and transmitting its conceptualization to the SD.
2. An object of consciousness is a momentary ID-object resulting from an effort to reflect an SOA in a conceptualization.
3. A desire is a lasting SD-object possibly causing an effort to reflect a conceptualization in an SOA.
4. An intention is a momentary ID-object possibly causing an effort of the intender to reflect a conceptualization in a future W-SOA and resulting from a desire with the same conceptualization.

These processes may not remain inside the human mind. The human can perform an action accordingly, physically, and/or through language. If it is the latter, the result of this expression will be reflected with modality and modal markers in a sentence.

Returning to the example of the ambiguous sentence 21, 'John may enter the room', it indicates the possibility that a person named *John* will end inside the space of *the room* (possibility modality). Furthermore, it can also be interpreted as a reflection of the knowledge of the speaker (epistemic modality) or a permission of the speaker (deontic modality). A SOA has been generated inside the mind of the speaker triggered by the information received from the world perceived, the OOM: the possibility that John entering the room is true. However, this can be realised in two ways: The mind of the speaker has interacted with a SOA involving John and the room. If, for example the speaker has seen John approaching the door of the room, the OOM, the section of his/her mind interacting with it, the consciousness, would have conceptualised the idea of John entering the room, and this belief would have travelled to the SD, eventually leading to the thought that it was possible for John to enter the room since he was about to.

The opposite process could have been also taken place: The SOA of John and the room could have been conceptualised in the speakers SD through a desire, and

would result in an intention generating a conceptualisation in a future W-SOA of John entering the room leading to a deontic modality in language. In this case, since the concept would be possibly true, that is, possible in some of the *worlds* of the speaker, the marker used to signal it in the sentence is *may*.

Taking necessity and possibility as the main points of the classification can be easily formalised into rules for the tagger. Firstly, since it is based on a logic perspective and it does not take a specific language as reference, it can be used in a cross-linguistic and, if necessary, in a cross-discourse study. Secondly, it can easily overcome the ambiguity problem between markers. Thirdly, the issue of negation can also be easily formalised. The main problems of the computational study, ambiguity and portability, could be resolved if we apply logic.

It may be incomplete, however, to select modality as necessity or possibility for several reasons. It would not offer us a comprehensive array of information of the intentions of the speaker, and we may leave out some modal markers. Moreover, the number of Japanese scholars following this view is a minority. Nevertheless, taking into account the variety of interpretations in this field, each study will be limited according to the point of view taken. This study is not trying to provide a definite answer to modality, but to consider the best approach for a computational study.

We now know what is modality and how it is going to be classified. The next discussion will cover how it is coded in a sentence. The next section will define what a *modal marker* is for this study, and move on to specifically talking about Spanish and Japanese, describing which markers we can find in each language and the problems that we may encounter by doing so.

2.3 Definition of modal marker

2.3.1 Establishing the analysis

Defining a modal marker can be as gruesome as defining modality. We have clarified up to this point that modality is a psychological process related to the necessity or possibility of an state of affairs by the speaker, which starts inside its mind, and realises through language in a sentence. The question to answer now is where to find this realisation in the sentence.

We have also clarified that this study is focused on tagging modality markers. These should encode modality overtly and, moreover, we want to move away from the speech act approach of modality, and focus on the grammatical markers. Which leads us to the question of verb mood. Section 2.1 explains that modality and mood are inevitably related. But is modality the same as mood? Is mood only marked morphologically in a verb? Should we include additional auxiliaries? Mood as we have seen is considered by many as the grammaticalisation of modality. However, consider the following sentences in Spanish:

(27) a. *El tren lleg-a mañana*
The train arrive-PRES tomorrow

'The train arrives tomorrow'

b. *El tren a lo mejor lleg-a mañana*
The train maybe.MODADV arrive-PRES tomorrow

'The train probably arrives tomorrow'

c. *El tren deb-e lleg-ar mañana*
The train must-PRES.MODAUX arrive-INF tomorrow

'The train must arrive tomorrow'

(28) a. *Esper-o que el tren lleg-ue mañana*
Hope-PRES that the train arrive-SBJV tomorrow

'I'm hope that the train arrives tomorrow'

b. *Depend-e de que el tren lleg-ue mañana*
Depend.PRES on that the train arrive-SBJV tomorrow

'(It) depends on the train arriving tomorrow'

c. *El tren probablemente lleg-ue mañana*
The train probably.MODADV arrive-SBJV tomorrow

'The train probably arrives tomorrow'

(29) *Ven mañana*
Come.MODIMP that

'Come tomorrow'

Sentences in 27 contain a verb in the indicative mood, sentences in 28 a verb in the subjunctive mood and 29 in the imperative. Do all sentences contain a modal marker? There answer is no, at least not for the purposes of this study.

Among the sentences with indicative, sentence 27a is a plain, declarative affirmative sentence. Also sentences 27b and 27c, but with a small exception: a modal marker has been added, an adverb in 27b, and auxiliary *deber* ('must, have to') in 27c. The indicative mood represents a fact of reality, or *realis*, (Moreno Cabrera, 2000), but it is not *overtly* marking the necessity of this reality becoming true, as does the adverb or the auxiliary.

A similar case can be found in 28 but with the subjunctive mood. The subjunctive represents the *irrealis*, the non-reality (Moreno Cabrera, 2000) or the mood of the subordination (Bosque, 2012, p. 378), but this is once again a very wide and vague description. The subjunctive on its own does not provide us sufficient

semantic content, and it cannot appear on its own. It is only used in combination with other elements of the sentence. They specify the meaning of the subjunctive, acting as 'triggers' or 'selectors of the grammatical mood', such as modal adverbs, as in Sentence 28c. Some modal adverbs may take the indicative mood like in 28b, others the subjunctive, some both (Bosque, 2012, p. 377). For this reason, these modal triggers will be the ones considered modal markers for this study.

The only mood that contains strong and sufficient modal sense is the imperative mood, as in sentence 29, signalling a necessity condition on the receiver of the message. The same situation can be found in Japanese, as the following sentences show:

(30) a. 車　　を　買-う
kuruma *wo* *ka-u*
car　　ACC buy-PLN

'(I) am buying / will / going to buy a car'

b. 多分　　車　　を　買-う
tabun　　*kuruma* *wo* *ka-u*
maybe.MODADV car　　ACC buy-PLN

'Maybe (I) will buy a car'

c. 車　　を　買-う-つもり　　だ
kuruma *wo* *ka-u-tsumori*　　*da*
car　　ACC buy-PLN-plan.MODADV COP

'I am going (intend) to buy a car'

(31) 車　　を　買-わ-なければならない
kuruma *wo* *ka-wa-nakerebanaranai*
car　　ACC buy-IRR-must.MODADV

'(I) have to buy a car'

(32) a. 車　　を　買え！
kuruma wo kae
car　　ACC buy.MODIMP

'Buy a car!'

b. 車　　を　買える
kuruma wo kaeru
car　　ACC buy.MODPOT

'I can buy a car'

As with the Spanish sentences, the ones in 30 use the basic form of the verb, equivalent to the indicative mood, but only sentences 30b and 30c contain a modal marker, in the form of an adjective (30b) or an auxiliary (30c).

The verb in sentence 31 is formed by a stem and an auxiliary. The stem is in the so-called *irrealis* or *negative* form, and is followed by the *nakerebanaranai* auxiliary that denotes a necessity. This form, one of the roughly equivalents to Spanish subjunctive, cannot appear on its own, as in Spanish. Its usage is *triggered* by another element, in this case a modal auxiliary, and without it its semantic content is empty and the sentence would be ungrammatical.

Finally, as with Spanish, the imperative mood is the only one considered as a modal marker on its own as it contains sufficient modal meaning (32a). In Japanese, the imperative inflection of the stem can stand out independently without making the sentence ungrammatical.The same case can be found in the potential mood of a verb, present in Sentence 32b. The verb, as with the imperative, is inflected in a form that contains sufficient modal information, in this case possibility, to be used on its own.

The answer of these differences among moods in both languages relies on the history of each language. As Jiménez Juliá (1989) explains, what we understand today as a modal marker in any European language has its origins in the Proto-

Indoeuropean's realisation of mood. In this proto-language there were at least five different types of mood: indicative, imperative, injunctive, subjunctive and optative. With the fragmentation into many different languages only three remained: indicative, imperative and subjuncitve mood, this last one formed by the union of the previous injunctive, optative and subjunctive moods. This unification lead to the creation of additional resources to express modality, such as auxiliary verbs, some of them replacing the grammatical mood like English modal verbs, others reinforced it, like adverbs in Spanish.

Therefore, in Spanish and Indoeuropean languages in general, the grammatical mood of the verb is the morphological manifestation of modality, which, in some languages such as English, has been partially replaced by other modal markers. However, Spanish mood needs additional auxiliaries to convey a specific modal meaning (with the exception of the imperative) which are responsible for the selection of the indicative or the subjunctive mood of the verb.

Japanese is not an Indoeuropean language, but has suffered a similar transformation. Old Japanese moods had sufficient content to stand on their own with inflectional suffixes, but in modern Japanese they need auxiliaries with a higher lexical value in order to be used (Iori, 2014, 45). The inflectional suffixes in Old Japanese were obligatory, and were used for syntactic, conjunctional and some modal categories such as the imperative, prohibitive or the desiderative. They could be preceded by optional auxiliaries that specified respect, aspect, voice, negation and *tense moods*: the *modal past* (marking a *hearsay, sudden realisation* or *emphasis*); the conjectural (probability, necessity, volition); or the subjunctive (counterfactual). The inflections could also be followed by extensions or clitics that added further necessitive, conjectural or evidential meanings (Frellesvig, 2010).

This changed in Middle Japanese, especially in the late period. The optional auxiliaries were all lost or turned and melted into inflectional suffixes like subjunctive or tense except for the negative, which remained marked. The originial inflectional suffixes were also lost, with the exception of the imperative. Other forms, mainly adjectives, began to grammaticalise themselves into new auxiliaries, similar to the forms we see today, and *filling in* the gap left by the loss of former auxiliaries

and inflectives, completing the process in Modern Japanese (Frellesvig, 2010). For example, subjunctive was ended with the inflection suffixes *mu, ramu, kemu, masi, zu*, and *besi*, such as in the sentences *Nanzi asu kurubesi* and *Asu ame hurubesi* ('You must come tomorrow', deontic reading, and 'It must rain tomorrow', epistemic) – from Onoe (2004), in Iori (2014, 46). These suffixes have faded away, and modern Japanese modal auxiliaries, grammaticalised forms of former lexical items such as adverbs (originated from adjectives, verbs, or Chinese loan words) or adjectives, are required to specify the meaning of the verb, as in *Kimi wa asu konakerebanaranai* ('You must come tomorrow', deontic modality) and *Asu wa ame ga furunichigainai* ('It must rain tomorrow', epistemic modality).

Hence, the answer to whether mood is a marker of modality is yes, but not quite, at least for the purposes of this study. If mood has sufficient content on its own like the imperative, it will be considered a modal marker, but otherwise, the additional elements added to the verb will be the ones tagged as markers. What is considered an additional element should be the next question to answer. The best answer for good productivity is, once again, simplicity. A modal marker has to fulfil three conditions:

1. Modality is something *marked* in language.
2. A modal marker must be recognised by previous works on the area, whether because it has been fully grammaticalised into an auxiliary, or its usage has been reduced exclusively to marking modality.
3. A modal marker must modify the sentence root (according to dependency grammar, the verb) or be the root in case of an adjective.

These criteria have been heavily influenced by Bybee et al.'s work (1994). They establish their object of study as grammatical morphemes that (1) must belong to a closed class, (2) must have a fixed position in relation to the verb, (3) must be lexically general and (4) must have predictable meaning in most contexts.

I understand markedness in a pair of elements as the one differentiated by the presence or absence of certain property (Ingram et al., 2016). More specifically, the unmarked form would be the plain, affirmative, indicative mood, and the marked

one any other which contains a modal marker. In other words, a modal marker is a *mark* (Gvozdanovic, 1989) added to the plain verb form that includes a modality meaning. Taking for example the following sentences, similar to the ones seen above:

(33) *Mañana com-o en casa*
Tomorrow eat-PRES at home

'Tomorrow I will eat at home'

(34) *Mañana com-eré en casa*
Tomorrow eat-FUT at home

'Tomorrow I will eat at home'

(35) *Mañana v-oy a com-er en casa*
Tomorrow go-PRES.MODAUX to.CONN eat-INF at home

'Tomorrow I will eat at home'

Sentences (33), (34) and (35) obtain the same meaning, and could be considered to signal necessity as in 'The fact of eating tomorrow at home is necessary/necessarily true'. However, this interpretation will lead to the tagging of all declarative sentences, which will be clearly unproductive for the purpose of this study. We will only consider sentence (35) to be *marked* and to have a modal marker (in this case, the periphrastic construction *ir a + V*, 'will V') that has been added to reinforce the indicative mood. Sentences (33) and (34) are affirmative sentences and only use the indicative mood, and will be considered unmarked. The same case can be found in Japanese. As Larm (2006) states, there are many ways modality can manifest itself, the most explicit being the one marked with the grammatical system. The plain infinitive verb or unmarked form may be used also for expressing the present tense, the continuous form, the imperative and also intentions made by the speaker (Horie & Narrog, 2014) (Sentence 36), roughly equivalent to the modal auxiliary つもり (*tsumori*, Sentence 37):

(36) 日本 に 帰-る
nihon *ni* *kae-ru*
Japan LOC go back-PLN

'I will go back to Japan'

(37) 日本 に 帰-る-つもり だ
nihon *ni* *kae-ru-tsumori* *da*
Japan LOC go back.PLN-plan.MODAUX COP

'I will (planning to) go back to Japan'

Grammaticalisation is "the change whereby lexical items and constructions come in certain linguistic contexts to serve grammatical functions, and once grammaticalized, continue to develop new grammatical functions" (Hopper & Traugott, 2003, p. 18). With grammaticalised modal markers we refer to those constructions, mainly verbs or adjectives, that have become modal auxiliaries or even suffixes (Traugott, 2006, p. 110), undergoing processes such as semantic generalisation, semantic reduction, bleaching, erosion, etc. (Bybee et al., 1994, p. 6). We will not, at least in this work, consider auxiliaries that are in the middle of the grammaticalisation process. For example, there are a number of constructions in Spanish in the process of becoming an auxiliary verb of a periphrastic construction, as the verb *querer*, 'to want', which has already been taken as a *semi-modal* (RAE, 2009), as represented in the following sentences.

(38) *Quier-o* *compr-ar* *un* *coche*
want-PRES buy-INF a car

'I want to buy a car'

(39) 車　　が　買-い-たい
kuruma ga ka-i-tai
car　　NOM buy-CONT-want.MODaux

'(I) want to buy a car'

Sentence (38) and Japanese equivalent (39) refer to the same thing. However, only the Japanese, (39), will be tagged with modality, as the Japanese suffix たい (*tai*) meaning 'desire' has already been fully grammaticalised into a suffix from a lexical adjective, whereas its Spanish equivalent still has not.

Following the grammaticalised auxiliaires, we will also consider for this study a series of adverbs as encoders of modality, such as 'probably' or 'maybe' and also predicative adjectives. Although they have not been grammaticalised as an auxiliary, their meaning is limited to probability values, their position and function is fixed, and they have been considered by previous works as modal markers (Spanish modal adverbs are labelled 'mood adverbs'). Also, for example, there is evidence that Old English already had a series of modal adverbs with modal meaning, that were used to emphasise the truth value or importance of a statement by the speaker, which were increased in Middle English with French borrowings and Early Modern and Modern English with the appearance and high usage of low, medium and high probability adverbs, many of which were originated in soon-to-be modal auxiliaries such as 'maybe' from 'may' –Swan (1991) in Traugott (2006). Since adverbs appear to have a relative importance in the development of modality in some languages, they will be included in the study along with a series of modal adjectives, for the same reason. (As we will see below, Spanish modal adverbs are formed from these adjectives). One of the questions that can answered from the quantitative study is their frequency of use in comparison to other fully grammaticalised markers.

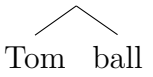
Finally, as before, we will be working with modality from a syntactic-semantic point of view. More specifically, a modal marker is the linguistic element that modifies the root of the sentence, or main verb, according to the rules of dependency grammar (DG). The reason we have chosen DG is because of its increased use in

computational and typological studies. Constituency grammar (CG) groups words in constituents of phrases which are organised hierarchically around a head. Dependency grammar, on the other hand, works with words individually, linked in pairs in head-dependent relations called dependencies. Citing Kubler et al. (2009, p. 2):

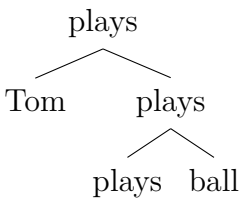
> The basic assumption underlying all varieties of dependency grammar is the idea that syntactic structure essentially consists of words linked by binary, asymmetrical relations called dependency relations (or dependencies for short). A dependency relation holds between a syntactically subordinate word, called the dependent, and another word on which it depends, called the head.

Both dependency and constituency grammars obtain the same thing, a syntactic structure of a group of words. However, one grammar can express things that the other cannot, and vice versa. The following trees represent the difference between DG (40a) and CG (40b) –taken from Osborne (2014, p. 604).

(40) a.



b.



Since it is concerned with the relations between words instead of grouping them, dependency grammar has proven very useful for syntactically parsing languages with free word order, like Japanese. Previous studies (Kurohashi & Nagao, 1994) have shown that Japanese's free word order and its variety of ellipsis possibilities are very difficult to handle with a CG parser, which suppose phrase order as specified in grammar rules. Dependency grammar can tackle very efficiently these problems, since head-dependent relations are not influenced by word order or ellipsis, as we can see in example 41 (taken from Kurohashi & Nagao (1994, p. 2):

(41)

It is necessary to clarify an issue regarding Japanese dependency grammar, related, once again, to the ever-lasting problem in Japanese parsing of *segmentation*. Although DG connects single words, for Japanese we should say it connects the *bunsetsu*[9] of a sentence. A bunsetsu is formed by one or more lexical words with zero or more functional words. For example, taking case particles as function words, a subject bunsetsu can be the noun denoting the subject + the nominative particle. Another example is verbal morphology. Tense, aspect, modal, politeness and other suffixes are considered as part as the same *bunsetsu* as the verb. A *bunsetsu* is then not as complex as a phrase in the Western sense, but if we consider functional elements like particles as *words*, they are bigger than independent words. The following example 42, from the study made by Murata et al. (2001), shows an example of a Japanese sentence, divided into *bunsetsu* by the dependency grammar:

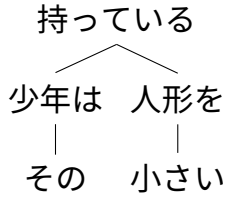
(42) a. その 少年 は 小さい 人形 を 持っ-て-いる

sono *shōnen* *wa* *chiisai* *ningyō* *wo* *mot-te-iru*

that boy NOM small doll ACC have-TE-be-PLN

'That boy has a small doll'

b.



Of course this is not the only and definite way of analysing words. Though it is the traditional and the one use by some automatic parsers, such as the KN Parser (KNP)[10] (Kurohashi & Nagao, 1994), we can keep dividing the words even further, adding for example a case dependency between a lexical word and its case particle (Asahara et al., 2002). The NINJAL[11] has proposed three forms of dividing the

[9] 文節, lit. 'section of a sentence'
[10] Kurohashi-Nagao Parser
[11] The National Institute for Japanese Language and Linguistics

words from a sentence (Tanaka et al., 2016, p. 1651):

1. Short Unit Word (SUW): SUW is a minimal language unit that has a morphological function. SUW almost always corresponds to an entry in traditional Japanese dictionaries. For example かもしれない (*kamoshirenai*, auxiliary denoting probility) is divided into three auxiliaries: かも (*kamo*), しれ (*shire*) and ない *nai*).
2. Middle Unit Word (MUW): MUW is based on the right-branching compound word construction and on phonological constructions, such as an accent phrase and/or sequential voicing.
3. Long Unit Word (LUW): LUW refers to the composition of bunsetsu units. An LUW has nearly the same content as functional words bounded by bunsetsu boundaries. In this case, the auxiliary かもしれない, for example, is treated as a whole.

The LUW is the traditional way of separating words, used by syntactic parsers (Kurohashi & Nagao, 1994, 1998; Kudo & Matsumoto, 2002) and treebanks and corpora such as the Kyoto University Text Corpus (Kurohashi & Nagao, 1998), the JPTenTen11 (Srdanovic et al., 2013) and the Balanced Corpus of Written Japanese (Maekawa et al., 2014). The SUW model is more recent, and is being used by the Universal Dependencies group (Tanaka et al., 2016) to achieve a typological tagging of dependencies (Nivre, 2015).

For this study, we will use the LUW separation for the tagging of modality, as the SUW may, in this case, gives too many information that will not be needed. Modal markers such as the previous かもしれない or なかれべればなりません (*nakerebanarimasen*, auxiliary denoting necessity, formal), although comprised by smaller auxiliaries, have been grammaticalised as a whole, and thus will be treated as a single entity that modify the main verb.

Although Spanish does allow an easier constituent analysis, we will also be using dependency grammar as reference for the selection of modal markers. The only issue that needs to be clarified is the role of the copula and the predicate in copulative sentences. The analysis in mind is the same followed in the Universal Dependencies

syntax, reflected in the Google Spanish corpus (Mcdonald et al., 2013) and the Ancora corpus (Talué et al., 2008) that consider the predicate as head or root of the sentence. However, other analysis such as the one performed by the Freeling Dependency Parser (Lloberes & Castellón, 2010) situate the copula as the root of the sentence. We have chosen the first solution (copula is not the head) because it appears to be the preferred one in dependency analyses, and it is the same analysis made for Japanese. In the following copulative sentences, the adjective *necessary* (*necesario* for Spanish, 必要 (*hitsuyō*) for Japanese), is a necessity modal marker. The analysis used for this study for Japanese and Spanish would be (43b.) and (44b.) respectively, as opposed to Freeling's (44c.):

(43) a. 食物 は 生命 に 必要 だ
shokuryō wa seimei ni hitsuyō da
food NOM life DAT necessary.MODADJ COP

'Food is necessary for life'

b. 必要だ
食物は 生命に

(44) a. *La comida es necesaria para vivir*
the food COP necessary.MODADJ for life

'Food is necessary for life'

b. necesaria
comida es vivir
La para

c. es
comida necesaria vivir
La para

According to these three requisites (they must be marked, grammaticalised or chosen by previous studies, and modify the sentence root or function as an adjective root), the modal markers selected for each language are the following:

1. Spanish: modal auxiliaries (periphrastic constructions), adverbs, adjectives, imperative mood, negative imperative.
2. Japanese: modal auxiliaries, adverbs, adjectives, imperative and potential moods.

Each one of them will be described below, starting first with grammatical moods, and followed by the auxiliaries, adverbs and adjectives in each language.

2.3.2 Departing from mood

The only mood that has sufficient content to mark modality is the imperative, used for a strong solicitation, as a necessity-deontic one, in both Spanish (45a) and Japanese (45c). This includes the Spanish negative imperative (45b), formed by a negative element followed by the subjunctive, indicating a prohibition, also a necessity-deontic modality. These constructions will only appear in independent lexical verbs. An auxiliary cannot accept an imperative (RAE, 2009, p. 800).

(45) a. *Ven* *mañana*
come.MODIMP tomorrow

'Come tomorrow'

b. *No* *veng-as* *mañana*
NEG come-MODSBJV tomorrow

'Do not come tomorrow'

c. 明日 来い
ashita *koi*
tomorrow come.MODIMP

'Come tomorrow'

Another morphological mood considered is the Japanese potential mood (46) or ”potential verbs” by some grammars (Kaiser et al., 2013, p. 398). It expresses a possibility-deontic/epistemic modality and the values of possibility and ability.

(46) コメント を 出せ-ない
komento *wo* *dase-nai*
comment ACC come out-NEG.MODPOT

‘(I) can’t comment’

2.3.3 Markers attached to main verb

The elements that reinforce the morphological mood are morphemes, either free or bound, that modify the main verb of the clause and add modal meaning. In Spanish these morphemes are auxiliary verbs, similar to English modal verbs, that together with the main verb form a multi-word sentence root called periphrastic construction. In the case of Japanese, due to its agglutinative nature, these morphemes are suffixes or auxiliaries attached to the main verb. For the sake of clarity, for the tagging of the corpora and the automatic implementation of the tagger, we will call all these forms ‘auxiliaries’.

2.3.3.1 Spanish auxiliaries

As explained further on, periphrastic constructions, or simply periphrases, are the most frequent way to code modality in Spanish. They are formed by two principal components: an auxiliary verb and the main verb, sometimes joined with a connective, with different characteristics, as shown in Table 4 (RAE, 2009, p. 529):

Both the auxiliary and the main verb form a single entity that acts as the head of the sentence(Gómez Manzano, 1991, p. 53). The auxiliary gives tense and modal information, and the main verb selects the subject and the complements These constructions are used to complement features of the inflection of the verb, or

Table 4: Composition and characteristics of Spanish periphrasis

Auxiliary verb	**Main verb**
- Does not select neither subject or complement	- Selects subject and complement
- Finite form: inflected for person, number, tense and mood	- Non-finite: infinitive, participle or gerund forms
- Closed list of grammaticalised verbs	- Open list of verbs

add new features that cannot be expressed morphologically, such as tense, aspect, modality and voice (Gómez Manzano, 1991, p. 82), and are divided into three groups according to the form of the main verb:

1. Infinitive periphrases
2. Participle periphrases
3. Gerund periphrases

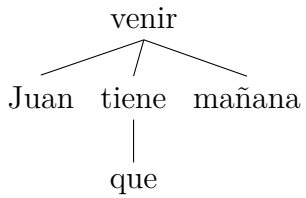
The constructions relevant to this study are the ones that express modality, which belong to the group of infinitive periphrastic constructions, that is, those which have the main verb in infinitive form. There are seven periphrases considered to convey meanings of necessity and possibility:

1. *Deber* + V ('must', 'have to')
2. *Deber de* + V ('must', 'have to')
3. *Haber de* + V ('must', 'have to')
4. *Haber que* + V ('must', 'have to')
5. *Tener que* + V ('must', 'have to')
6. *Poder* + V ('can', 'may')
7. *Ir a* + V ('will', 'going to')

Even though the modal meaning is assigned by the auxiliary, since it forms a multi-word head with the main verb, we shall consider both of them (along with the connective, if any), the modal marker of the sentence. As shown in 47a.:

(47) a. *Juan tien-e que ven-ir mañana*
Juan must-PRES.MODAUX to.CONN come-INF tomorrow

'Juan must come tomorrow' (The fact 'Juan coming tomorrow' is necessary)

b.


Dependency syntax does not portray a periphrastic construction as a single unit (see 47b.). Instead, it takes the main verb as the head of the sentence and the auxiliary as a modifier. However, this does not change our analysis. In fact, we strongly believe these should be analysed together in dependency grammar, although this discussion may be continued in future works. The nature and grammaticalisation of an auxiliary obliges it to be followed by the main verb it modifies.

2.3.3.2 Japanese auxiliaries

A similar situation can be found in Japanese. One of the most frequent modal markers are suffixes and auxiliaries carrying the modal meaning that are attached to the main verb. The liaison is nearly identical to the Spanish periphrases, which reinforces our claim for the analysis.

Table 5: Composition and characteristics of Japanese auxiliaries

Auxiliary	**Main verb**
- Does not select neither subject or complement	- Selects subject and complement
- Finite form: inflected for tense, mood and negation	- Non-finite: irrealis, adverbial, conclusive, attributive, hypothetical stems, te-form or ta-form
- Closed list of grammaticalised elements	- Open list of verbs

The main difference from Spanish periphrases has to do with the agglutinative nature of Japanese. Whereas in Spanish it is a combination of two or more words, in Japanese the suffixes or auxiliaries are attached directly to the inflected form or stem of the main verb, more similar to the Turkish-type than the inflectional morphology of the Latin-type (Shibatani, 1990, p. 221). A Japanese verb is formed by a stem to which auxiliary affixes are attached to carrying information such as tense, aspect, modality, politeness, etc.

There are many denominations to the parts of the verbs, mainly due to the translation to English. Matsuoka McClain (1981) and Nomura (2010) name the verb stem as 'base', which is formed by a stem and a 'base formative' followed by suffixes, as illustrated in Figure 7.

Figure 7: Matsuoka McClain and Nomura's Japanese verbal composition

$$\text{Japanese Verb} = \underbrace{\text{Stem} + \text{Formative}}_{\text{Base}} \ (+ \text{ Auxiliary}) \mid (+ \text{ Particle})$$

Other authors such as Shibatani (1990) name the base as 'inflectional category' or 'stem', which is formed by the root and an inflectional ending, and the following elements are auxiliaries (see Figure 8). Iwasaki (2013) follows a similar labeling, naming the base as 'Stem', formed by a Root and a 'Stem Forming Suffix', followed by auxiliary suffixes. For this study, as well as Spanish, for the sake of clarity, I will follow Shibatani's labeling for the description and consider them as auxiliaries, whether or not they could be considered suffixes, attached to the stem.

Figure 8: Shibatani's Japanese verbal composition

$$\text{Japanese Verb} = \underbrace{\text{Root} + \text{Inflectional ending}}_{\text{Stem}} \ (+ \text{ Auxiliary})$$

The most important feature of this composition is that each stem subcategorises the following auxiliaries. For example, the desiderative suffix たい (*tai*) can only be attached to the *continuative* stem of the verb. The verb 'to see' (みる, *miru*) has the

continuative stem み (*mi*). The verb form 'want to see' is formed by joining both elements: み-たい (*mi-tai*). The suffix *tai* cannot join the verb if it is not inflected in its continuative form. The list of all modern Japanese stems using the verb 'to see' as an example and some auxiliaries is illustrated in Table 6.

Table 6: Japanese inflection stems

Stem	**Root**	**Inf. Ending**	**Auxiliaries**
未然形 (Irrealis or Negative)	*mi*		*nai* (Negative), *rareru*, *sareru* (Voice), etc.
連用形 (Adverbial or Continuative)	*mi*		*masu* (Polite), *tai* (Desiderative), *soo* (Conjectural), etc.
終止形/連体形 (Conclusive, Attributive)	*mi*	*ru*	*deshō*, *rashii* (Hearsay), Infinitive form, etc.
仮定形 (Hypothetical or Conditional)	*mi*	*re*	*ba* (Conditional), etc.
命令形 (Imperative)	*mi*	*ro*|*yo*	
テ形 (te-Form)	*mi*	*te*|*yo*	
タ形 (ta-Form[12])	*mi*	*ta*|*yo*	

Modal auxiliaries belong to the 'Auxiliaries' column of Table 6, such as desiderative *tai* previously mentioned. They are morphemes that are attached to a specific stem. The subcategorisation is important for this study, as modal auxiliaries will not combine with every stem, and so must be specified in the tagger.

In contrast to Spanish list of 7 auxiliaries, in Japanese there are at least 23, without counting their formality or phonetic variations:

1. V + なければならない (*nakerebanaranai* Have to / Must)
2. V + ざるを得ない (*zaruwoenai* Have to / Must)
3. V + しかない (*shikanai* Have to / Must)
4. V + 訳にはいかない (*wakenihaikanai* Cannot / Must not)

5. V + に忍びない (*nishinobinai* Cannot / Could not)
6. V + べき (*beki* Should / Ought to)
7. V + た方がいい (*tahōgaii* Should / Ought to)
8. V + たらいい (*taraii* Should / Ought to)
9. V + ればいい (*rebaii* Should / Ought to)
10. V + たい (*tai*))
11. V + てもらいたい (*temoraitai*))
12. V + ほしい (*hoshii*))
13. V + ください (*kudasai*))
14. V + つもり (*tsumori* Will / Going to)
15. V + かねる (*kaneru* Cannot / Could not)
16. V + はず (*hazu* Will / Going to)
17. V + に違いない (*nichigainai* Will / Going to)
18. V + てもいい (*temoii* Can / May)
19. V + ことができる (*kotogadekiru* Can / May)
20. V + かもしれない (*nishinobinai* Can / May)
21. V + とは限らない (*tohakagiranai* Not have to)
22. V + ほどのこともない (*hodonokotomonai* Not have to)
23. V + だろう (*darō* Can / May)

In conclusion, both Spanish and Japanese have modal auxiliaries that provide the main verb the modal meaning. In Spanish we have independent words that precede the main verb forming a multi-word sentence root. In Japanese the auxiliaries are attached directly to the main verb that is inflected in a specific stem. The array of auxiliaries is much higher in Japanese, which means that their semantic content will be much more limited and specific, and hence, the ambiguity will be lower than in Spanish. These constructions will be the most frequent way of coding modality in both languages, as seen below. The next section will describe another marker: modal adverbs.

2.3.4 Modal adverbs

Modal adverbs behave surprisingly similar in both languages, and therefore the explanation does not need two separate sections. They are a specific type of adverbs (content, non-inflected words that modify verbs, adjectives and other adverbs) that provide necessity or possibility meanings to the modified element.

Except for a few exceptions where they are independently created, Spanish modal adverbs usually are formed by adding the suffix *-mente* to an adjective. These adjectives must be descriptive, except those that express physic, state, spatial or temporal meanings (Rodríguez Ramalle, 2003, p. 12). Suffix *-mente* does not add new lexical value to the adverb, and, therefore, its modal content is established by the adjective. An example of the transformation is shown in Figure 48.

(48)

Necesario $\xrightarrow{+\ \textit{-mente}}$ *Necesariamente*

'Necessary' (Adjective) 'Necessarily' (Adverb)

Japanese modal adverbs can be formed through three different processes: (1) the adverb exists on its own, not formed by a derivation process (example 49); (2) it is derived by adding the suffix *ku* to an adjective (in a similar way to Spanish, see example 50); or (3) it is formed by an adjective followed by the particle に (*ni*) (see example 51).

(49) 多分

tabun

'Probably'

(50)

早い (*hayai*) $\xrightarrow{+\text{く}\ (ku)}$ 早く *hayaku*

'Rapid' (Adjective) 'Rapidly' (Adverb)

(51)
静か (*shizuka*)	$\xrightarrow{+に\ (ni)}$	静かに *shizukani*
'Calm' (Adjective)		'Calmly' (Adverb)

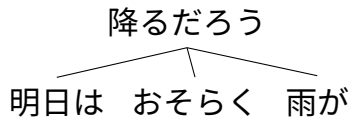
In both Spanish and Japanese, some adverbs may code the necessity meaning, while others the possibility one. However, all of them are considered epistemic modal markers. All of them are used to address the probability of an event by the speaker, but cannot be used to influence the hearer. For this reason, it is believed that their semantic scope modifies the whole sentence (Kaul de Marlangeon, 2002; Rodríguez Ramalle, 2003; Narrog, 2009a). In relation to dependency syntax, they appear modifying the main verb:

(52) a.

明日	は	おそらく	雨	が	降-る	だろう
ashita	*wa*	*osoraku*	*ame*	*ga*	*fu-ru*	*darō*
tomorrow	NOM	probably.MODADV	rain	NOM	fall-PLN	seem.MODAUX

'It will probably rain tomorrow'

b.

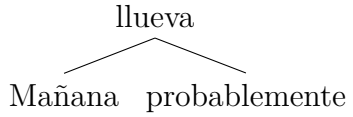


(53) a.

Mañana	*probablemente*	*lluev-a*
tomorrow	probably	rain- SBJV

'It will probably rain tomorrow'

b.



2.3.5 Modal adjectives

Another marker to be considered for the coding of modality is the adjective in a predicative position of a copulative sentence. As with the adverbs, the difference between Spanish and Japanese is quite small except for a few clarifications on the labeling.

In a copulative sentence, the main elements of the predicate are the copula and the predicative expression, which can be an adjective or nominal phrase. Since the copular verb contains little to no meaning and the predicative expression provides the subject with a specific characteristic, it is considered by many grammars as the most important part of the sentence, and hence, the head. As explained in Section 2.3.1, this will be the stance taken in this study for Spanish, because it is done as such in the Universal Dependencies, and Japanese, since the syntactic importance of the copula is too little. In Japanese, the copula will almost entirely be reduced as a signifier of tense, aspect and politeness to the sentence with no contentful meaning. In its informal, present form (だ, *da*), it can be omitted in the spoken language.

In Spanish, the main copular verbs are *ser*, *estar* ('to be') and *parecer* ('to seem'). These behave differently, with different or equal meanings depending on the predicate that follows. If it is an adjective, the overlapping of their meanings depends on the nature of the adjective, such as with *Es/Parece/Está frío* ('It is/seems cold') (Camacho, 2012, p. 455). With modal adjectives, however, verb *estar* appears to be incompatible with them, as in *Es/Parece/*Está posible* ('It is/seems possible'). According to Fernández Leborans (1995) (as seen in Camacho (2012, p. 456), adjectives denoting properties inherent to a genus or a species cannot be used with this verb. Also, these adjectives can receive the suffix *-mente* and become the previously mentioned modal adverbs. In other words, modal adverbs can only be used with contentful verbs, and adjectives with Spanish copula *ser* and Japanese *da*.

As in numerous occasions, there are many translations into English from Japanese for labeling this kind of adjectives, named 形容動詞 (*keiyō-dōshi*, 'adjective verb'). The most common label among grammarians is 'adjectival noun' or 'nominal ad-

jective', as they are in between *true* adjectives and nouns, as Shibatani (1990, p. 216) explains. As adjectives, they cannot possess grammatical functions of subject, object, etc. and cannot take nominative or accusative case particles. They can also be nominalised like any adjective with suffix *-sa*, unlike nouns. On the other hand, like verbs, *true* adjectives have tense and politeness inflections. Nouns and nominal adjectives do not, and require the copula for tense and politeness information. Also, some words can be used as a noun or as a nominal adjective. For example, 健康 (*kenkō*) can be read as 'healthy' or 'health'.

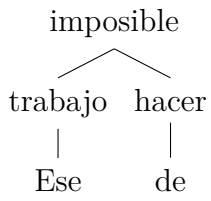
When modifying a noun, these adjectives will receive as suffix the copula in its attributive form (*-na*). For this reason, they are also commonly regarded as '-na adjectives' (Kaiser et al., 2013, p. 156), in opposition to the '-i adjectives' or *true* adjectives, since these always finish with the vocal 'i' in their plain form. Following these characteristics, other authors treat them as nouns, with names as 'adjectival noun' (Shibatani, 1990; Martín, 2004) or 'copular noun' (Nomura, 2010).

We cannot find a common treatment among the most widespread Japanese morphological taggers either. Juman will tag them as 'adjective' and the special tag '-na adjective', MeCab will simply use the 'noun' tag and ChaSen (Matsumoto et al., 2002) uses the 'noun' tag followed by the '-na adjective' subtag.. Therefore, since there is not a common approach to this case, we will refer to them as 'predicative adjectives', and tag them later as an adjective modal marker.

Sentences 54 and 55 offer examples of the usage of these adjectives encoding modality.

(54) a. *Ese trabajo es imposible de hac-er*
that job COP impossible.MODADJ to do-INF

'That job is impossible to do (for me)'

b.


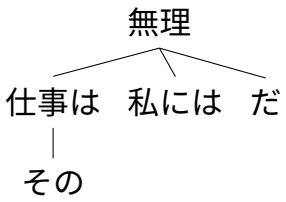
(55) a. その 仕事 は 私 に は 無理 だ

sono *shigoto* *wa* *watashi* *ni* *wa* *muri* *da*

that job NOM me ACC NOM impossible.MODADJ COP

'That job is impossible for me to do'

b.



To sum up, we have several ways of coding modality in Spanish and Japanese, not very different from each other: elements that are attached to a main verb, either with auxiliary elements or suffixes, adverbs and adjectives, as illustrated in Table 7 (influenced by Horie & Narrog (2014, p. 120), although we do not consider evidentials and discourse markers for our study).

Table 7: Modal System in Spanish and Japanese.

Spanish	**Japanese**
Auxiliaries - Periphrastic constructions with main verb (*poder, deber, tener*, etc.)	Auxiliaries - Attached as morphemes to the main verb (*-tai, -beki, -rashii, -kamoshirenat*, etc.)
Adverbs - Single and multi-words (*probablemente, necesariamente, a lo mejor*, etc.)	Adverbs - Single words (*zettai, tabun*, etc.)
Predicative adjectives - Preceded by copula (*posible, necesario*, etc.)	Predicative adjectives - Followed by copula (*muri, kanō*, etc.)
Imperative mood (*ven, no comas*, etc.)	Imperative mood and potential mood (*ike, mieru*, etc.)

2.3.6 Negation of modal markers

For those who consider modality as an expression of the subjectivity or the attitude of the speaker, negation can be understood as another marker that modifies the proposition (Masuoka & Takubo, 1992; Sanz Alonso, 1996). In this study, however, negation is not considered to carry modal meaning, but a semantic modifier of modality, an element attached to an auxiliary or a modal adjective that can change the type of modality it expresses, as we saw in Section 2.2.1. The most relevant issue concerns the modal auxiliaries: as they are formed by a pair of auxiliary verb and main verb, the negative element can affect either of the two. If the negation affects the auxiliary, modality changes, as indicated by rules of logic. If it affects the main verb, modality does not change. Sentences in 56 and 57 show this matter.

(56) a. *No pued-o com-er nada más*
NEG can-PRES.MODAUX eat-INF anything more

'I can't eat anything else' (= It is *not possible*, i.e. *necessary not*, to eat more)

b. *No deb-o com-er nada más*
NEG must-PRES.MODAUX eat-INF anything more

I mustn't eat anything else (= *necessary not*, to eat more)

(57) a. 賛成 が できｰません
sansei ga deki-masen
agree NOM can-NEG.MODNEG

'(I) can't agree (with you)' (= It is *not possible*, i.e. *necessary not*, to eat more)

b. 私 もう 当たり-た-くない から 最初 から
watashi *mō* *atari-ta-kunai* *kara* *saisho* *kara*
I again get hit-want-NEG.MODNEG since beginning from
外野 行く タイプ、
gaiya *iku* *taipu*
outfield go-PLN TYPE

'I don't want to get hit again so I'm staying in the outfield from the beginning' (= it is *necessary not* to hit)

Studies have shown that this occurs in many other languages (Palmer, 2001; Radden, 2014) and that overall it is a challenging feature for recent natural language processing tasks (Dowty, 1994; Wilson et al., 2009; Councill et al., 2010), and Spanish and Japanese are no exception. Although negation can modify modal auxiliaries, it does not affect them equally. As Kataoka (2012) defends, the issue relates to not only the scope but also the *polar* point of the negative element (Fauconnier, 1975; Ladusaw, 1979; Huddleston & Pullum, 2002). That is, although every construction formed by a main verb and an auxiliary or suffix is under the scope of the negative element, the polar point depends on the modal auxiliary used. In sentences 56a and 57a the polarity of negation is on the auxiliary. The modality is negated and hence, its type is changed, in this case, from a possibility to a necessity[13]. However, in other constructions the modality does not change, such as in 56b and 57b. The necessity is maintained, apparently breaking the rules of negative logic operations (Hintikka, 2002). In these cases, it is understood that the focus of the negation is on the main verb, and therefore not affecting the modality semantics. The discussion regarding which modal marker belongs to each type will take place in Chapter 3.

Grammatically, negation can be performed in many different ways, either at the lexical level, normally with an adverb or an auxiliary, at the morphological level with affixes, or semantically, with predicates expressing doubt, opposition, etc. (Bosque,

[13]Recall negation in logic, Example 15

1980, p. 26). These elements act as syntactic operators (RAE, 2009, p. 3631) that apply the negative notion to the constituents under their scopes or areas of affect or influence. We are interested on those lexical negative elements of the sentence that modify, or have a scope over, the modal marker, either Spanish or Japanese.

In Spanish, lexical negation is performed by different syntactic classes, mainly adverbs *no* ("no"), *nunca* ("never"), *jamás* ("never") and *tampoco* ("neither", "nor"). The problem when automatically processing modality is the correct detection of the negative element that affects the modal marker. The case of Spanish proves to be problematic especially in spoken discourse due to the separation of words: the negative element may appear in different positions of the sentence, and the modal marker can fall outside its scope.

In Japanese, negation of a predicative element is performed mainly through an inflection auxiliary, especially the grammaticalised adjective ない (*nai*) (Kaiser et al., 2013, p. 154). This particle is attached to those elements that can be inflective, i.e. verbs and adjectives. Predicative adjectives, since they are not inflected, must use the copula in its inflected negated form. Consequently, the negative ない (*nai*) has different variations. When attached to the copula, it becomes ではない (*de-hanai*) in formal contexts, but in more spontaneous ones it can also appear as はない (*hanai*), がない (*ganai*), じゃない (*janai*), りゃない (*ryanai*), etc. depending on the consonant of the preceeding syllable (Kaiser et al., 2013, p. 444). There are also variations depending on the type of discourse: in the written form it can appear as ぬ (*nu*), ず (*zu*), ざる (*nai*) or にあらず (*nai*), but in spoken language the ending *nai* can be shortened into ん (*n*). Finally, the formal equivalent of *nai* is the inflection ません (*masen*), turn into でわ/じゃありません (*dewa/jaarimasen*) if used with the copula. The Spanish problem of negation distance will not be as problematic in this language since the negative markers appear attached to the auxiliary. The main problem when processing the negation in Japanese is variation.

This concludes Chapter 2 of the study, where the theoretical ideas that will serve as the foundation for this study have been setted. First, we have made a brief overview of the history of the concept of modality and how it has been studied by the most important linguists, philosophers and psychologists in the last centuries,

and how it is considered today.

Secondly, we have explained the stance taken regarding modality, mainly defining it as a psychological connection between the mind of the speaker and a state of affairs (SOA) of the external world. The words that encode modality in the sentence will state if a SOA is necessary true or on the other hand possibly true. A second level of classification expresses an epistemic modality if the speaker *believes* the SOA to be necessary or possible, or a deontic one if he or she *desires* to be necessarily/possibly true. At this level, however, we may encounter a high amount of ambiguity since one marker may contain both epistemic or deontic readings.

Finally, we have clarified that modality is represented grammatically in a modal marker, an element of the sentence that is not present in all sentences, but only on those in which modality is overtly expressed. A modal marker needs to be *marked*, grammaticalised or registered by previous studies as an element with strong modal content, and has to modify a verb, according to the dependency rules of dependency grammar, as it is the most appropriate position for a comparative and computational study. The definition of a modal marker is subdued by the definition of modality. Here we have considered a more restricted approach, considering only those elements that modify a verb adding a necessity or a possibility meaning, completing the semantics of the verb mood.

The next chapter will depart from these ideas and explain the methodology followed by this study: the corpora and computational tools used, and how the tagging of the markers in the corpora was made, including the selected tagset.

Chapter 3

Methodology

3.1 Steps

The previous chapter has described the theoretical implications that will serve as foundation of the study. We will now move on to the description of the methodology along with the data and the tools used. Recalling the objective of the work, it is divided in three main parts:

1. Selection of the appropriate approach towards modality for this work.
2. Development of a quantitative comparable study of modality from Spanish and Japanese spoken corpora.
3. Automatic implementation of modality annotation for future studies

Chapter 2 has covered (1), the trends regarding modality from the last centuries until today. It concluded that modality signals the possibility or necessity of an state of affairs becoming true, whether perceived (epistemic) or desired (deontic) by the speaker. Modality is a semantic value, coded into grammatical elements that modify the verb, according to dependency syntax, which add specialised meaning to its mood. This Chapter 3 will describe the development process for (2) and (3), the following Chapter 4 will cover (2) and the final Chapter 5 will explain point (3).

We now shift from theoretical to empirical information, from the present to the future, how previous theoretical insights may apply to today's language and possible future texts. As with the theoretical decisions, the aim in this study is to follow a simple but precise and well-planned methodology: preparation, annotation, observation and implementation.

The preparation phase consists on two main steps: firstly, the configuration of the tagset that will be used in the annotation of the corpora and for the automatic tagger. The objective is a comparative annotation and study; hence, the procedure will use the same XML tags, symbols assigning descriptive information to elements of the text (Leech & Smith, 1999), for both languages. Secondly, a listing of each possible modal marker in both languages. It includes information found in the literature and personal knowledge, and recursively improved after observing the usage in the corpora.

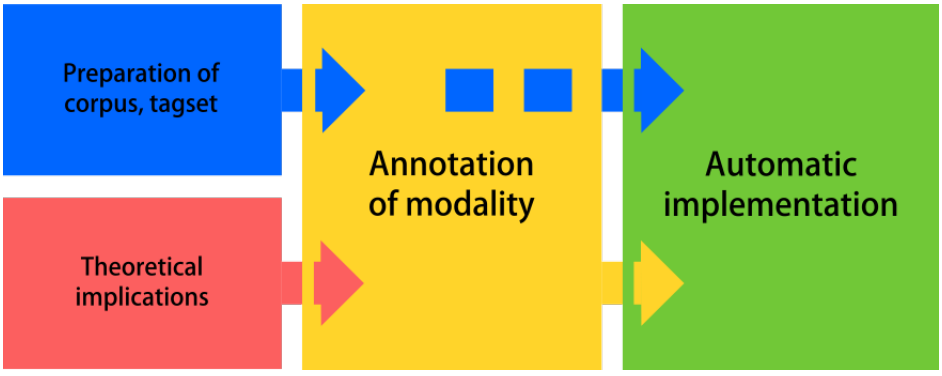
Next, in the annotation phase, after cleaning and preparing the corpora for the XML annotation, the markers compiled in the preparation phase were searched and tagged in the texts. The procedure was made manually and semi-automatically, using the established tagset. Any new information gathered from the text, such as new or different markers or problematic cases, was included in the listing. In our case, the XML tags assign modal information, as well as additional characteristics regarding the nature of the marker in the text: if it is negated or not, if it has an element missing through ellipsis or overlapping, if it is separated by other words, and if there is a misspelling or an error.

Following this, the observation phase takes place through a quantitative analysis of the modal markers found in the corpora. The objective here is to observe the usage of modality in a natural, spoken Spanish and Japanese discourse and confirm a series of hypotheses drawn. The results of these analyses will be presented in Chapter 4.

The last stage of the study is the automatic implementation of modality. That is, to develop a program that could automatically find and tag these markers in a new given text. The program is rule-based, and based on the theoretical information explained in Chapter 2 and the information extracted from the corpora study in Chapter 4. Its development, along as the problems and challenges that Spanish and Japanese present in this area, is explained in Chapter 5. The complete script of the program can be found in Appendix B.

Figure 9 summarises the steps taken in the study:

Figure 9: Methodology followed in the study



Each of step will be explained in the following sections: description and preparation of the corpora (Section 3.2), the annotation language and tagset (Section 3.3.2), the computational tools used along the study (Section 3.4) and the discussion of each modal marker (3.5).

3.2 Corpora

Two spoken corpora were used for this study, one Spanish (Spanish C-ORAL-ROM[1]), and one Japanese (C-ORAL-JAPÓN[2]), compiled and created by the Laboratory of Computational Linguistics of the Autonomous University of Madrid[3]. They were chosen for this study for two main reasons. First, both corpora have been created following the same procedure, making them very similar: a series of spontaneous, non-scripted conversations and monologues from real life that were recorded and later transcribed. The speakers in each corpus are native speakers from both sexes, different ages, parts of the country and education. There are, however, some differences related to the purpose of the compilation of each corpora, which forces us to avoid calling them *comparable* corpora: whereas C-ORAL-ROM was recorded as a general sampling of the language in mind, selecting as many situations and discourses as possible, without a time limitation for each speaker, C-ORAL-JAPÓN was created with the purpose of aiding with the teaching of the language. This reduced the variety of the topics of each recording, which were pre-established as different educational thematic situations. Also, the length of each recordings was set to be nearly the same.

Secondly, the spoken discourse was selected because previous studies have shown that modal markers are more frequent in spoken rather than written discourse (Gómez Manzano, 1991; Biber et al., 1999; Herrero & Moreno, 2014). It is believed that the usage of a natural context will provide more examples and possible exceptions to the rules.

The Spanish C-ORAL-ROM corpus (Moreno et al., 2005) has a total of 301,329 words and 379 different speakers (225 men and 154 women). It is divided into monologues (one speaker), dialogues (two speakers) and conversations (more than two), both formal and informal, from private, public, media and telephone contexts. The C-ORAL-JAPÓN corpus (Garrote et al., 2015) has 127,676 words and 58 (21 men and 37 women) speakers. It is also divided into monologues, dialogues and

[1]http://www.lllf.uam.es/ING/Coralrom.html
[2]http://www.lllf.uam.es/ING/Coraljp.html
[3]http://www.lllf.uam.es/ING/index.html

conversations in various situations. Table 8 shows an example of a header and a portion of the transcription from each corpus.

The corpora were cleaned and prepared before tagging the modality. Although each corpus contains numerous extralinguistic information such as overlapping, repetitions, reformulations, alternative writing of words (Japanese corpus), laughter, coughing, etc., the unnecessary marking for the present study was deleted for the sake of clarity, leaving only the raw text. The text from repetitions and reformulations was left behind. Table 9 shows an example of the corpora before and after the preparation process. The next section will describe the tools and mark-up language used for annotating these texts, as well as the computational tools used for processing the data and developing the tagger.

Table 8: Examples of header and text of the corpora already prepared for the tagging process

C-ORAL-ROM	**C-ORAL-JAPÓN**
<Header> <Title>yo era peleadora </Title> <File>efammn01 </File> <Participants> <Speaker> <Name>MAR, Marta</Name> <Sex Type="woman"/> <Age Type="D"/> <Education Type="1"/> <Occupation>retired</Occupation> <Role>participant</Role> <Origin>Chile/30 years in Madrid</Origin> </Speaker> </Participants> <Date>20/03/2001</Date> <Place>Madrid </Place> <Situation>at MAR´s living-room, hidden, researcher observer</Situation> <Topic>memories of childhood and youth</Topic> <Source>C-ORAL-ROM </Source> <Class Type1="informal" Type2="family-private" Type3="monologue '/> <Length>33' 41" </Length> <Words>4597 </Words> <Acoustic_quality Type="A"/> <Transcriber>Guillermo </Transcriber> <Revisor>Ana; Ana and Inma (prosody) </Revisor> <Comments> </Comments> </Header>	<Header> <Title>Japanese teaching</Title> <File>jmn01</File> <Participants> <Speaker> <ShortName>HOS</ShortName> <Sex Type="man"/> <Age Type="C"/> <Education Type="3"/> <Ocupation>teacher</Ocupation> <Role>participant</Role> <Origin>Tokyo</Origin> </Speaker> </Participants> <Date>28/07/2004</Date> <Place>Madrid</Place> <Situation>classroom</Situation> <Topic>how to teach Japanese through culture</Topic> <Source>CORALJAPANESE</Source> <Class Type1="formal" Type2="public" Type3="monologue"/> <Length>11'18"</Length> <Characters>3.879</Characters> <Acoustic_quality Type="B"/> <Transcriber>C. Kimura</Transcriber> <Revisor>K. Matsui</Revisor> <Contents/> </Header>
<Utterance id="8975" Type="interrogation">que éramos cuatro mujeres como te digo y cuatro hombres ya yo tenía a mi hermano para mí entonces yo tenía que hacerle todas sus cosas lavarle calcetines la ropa todo darle todo limpio y él me tenía que vestir a mí y darme lo que lo que yo necesitaba ves</Utterance> <Utterance id="8976" Type="suspension">libros ir a dejarme al colegio porque estaba un poquito lejos</Utterance>	<UNIT id="13600" speaker="HOS"> unkn 行ったり来たりというええそれについて少し考えてええみたいと思います。 </UNIT> <UNIT id="13601" speaker="HOS"> であのう <utterance/> ううんと先日あのう rep 昨日ですねあのええはっぴょ rep いろんなあの方々のはっぴょ rep あの非常に精力的な発表を伺って <utterance/> で僕もずいぶんあのう刺激を受けました。 </UNIT>

Table 9: Examples of corpora before and after the cleaning process

	Original corpus	**Corpus prepared for tagging**
C-ORAL-ROM	<Turn><Name>ROS</Name> <Says>**es que no sé nada** <Tone_Unit Type="standard" />**Patricia**<Tone_Unit Type= "standard" />**tía** <Utterance Type= "enunciation" />**si todavía no sabemos** <Utterance Type= 'interruption" /> **solamente yo sé**<Tone_Unit Type="standard" />**que yo** <Fragment>traba</Fragment> <Tone_Unit Type="total_restart" /> **yo en teoría no trabajaba** <Tone_Unit Type="standard" />**el** <Tone_Unit Type="standard" /> **jueves viernes sábado domingo antes**<Tone_Unit Type="standard" />**pero** <Tone_Unit Type="standard" /> **trabajo**<Tone_Unit Type="standard" />**entonces** <Tone_Unit Type="standard" />**estoy en Granada** <Utterance Type= 'enunciation" /> **y si estoy en Granada trabajando**<Tone_Unit Type="standard" />**dando clases en un pueblo que se llama Dúrcal**<Tone_Unit Type="standard" />**como la Rocío Dúrcal** <Utterance Type= 'suspension" /> <Overlap><Non_Linguistic /></Overlap> <Utterance Type= 'enunciation" /> <Notes Type= 'act">(54) laugh </Notes></Says></Turn>	<Turn> <Name>ROS</Name> <Says><Utterance id="131" Type= 'suspension">es que no sé nada Patricia tía si todavía no sabemos solamente yo sé que yo yo en teoría no trabajaba el jueves viernes sábado domingo antes pero trabajo entonces estoy en Granada y si estoy en Granada trabajando dando clases en un pueblo que se llama Dúrcal como la Rocío Dúrcal</Utterance> </Says> </Turn>
C-ORAL-JAPÓN	<UNIT speaker="HID" startTime="7.706" endTime="10.831"> **なんだか** & お [/] **手先がよく動くんですよ** ///</UNIT> <UNIT speaker="HID" startTime="10.831" endTime="30.799"> **それで** {%alt: そいで} # & す [/] **すごくねえ趣味が** -> / **いっぱいあったんだけど** / **その中でもね** // **姉様人形と手まりが好きで** // **どうしてこうなるのかなあってことを** / **考えて** // **で今** / **だいたい新聞広告見て** // **あのう** / **講習があると** [///] **やっぱお金使えないからね** ///</UNIT>	<UNIT id="13658" speaker= "HID"> なんだかお rep 手先がよく動くんですよ。</UNIT> <UNIT id="13659" speaker="HID"> それです rep すごくねえ趣味がいっぱいあったんだけどその中でもね <utterance/> 姉様人形と手まりが好きで <utterance/> どうしてこうなるのかなあってことを考えて <utterance/> で今だいたい新聞広告見て <utterance/> あのう講習があると <reset/> やっぱお金使えないからね。 </UNIT>

3.3 Annotation of modality

Three main resources where used for the development of this study: a markup language (XML), two part-of-speech taggers (Grampal and Juman) and a programming language (Python). The next subsections will describe each of them.

3.3.1 Using XML

XML (eXtensive Markup Language) was used for annotating modality in the corpora. XML is a type of *mark-up* language, which adds information to a text and organises it into a specific format. The information is added through *tags*, normally beginning and ended with the '<' and '>' symbols, determined and created by the user. In Corpus Linguistics, XML can be used for two purposes: corpus *mark-up* and corpus *annotation*. Mark-up refers to the system of codes that provide objective information about the text, its contextual information, type, genre, sociolinguistic variables, speaker information in spoken corpora, etc. (McEnery et al., 2006). An example of mark-up can be seen in the headers of our corpora in Table 8.

Corpus annotation refers to the process of adding interpretative, linguistic information to an electronic corpus (Leech, 1997, p. 2). That is, adding information to the text that can be later used for linguistic studies. While the mark-up is objective information that organises the text, annotation is interpretative, based on human knowledge. As McEnery et al. (2006, p. 30) explain, annotating a text provides several advantages: it makes it easier to extract linguistic information from the text, it is reusable and multifunctional, and it provides an explicit linguistic analysis and a reference resource. The most common annotations are part-of-speech (POS) tagging, lemmatisation and syntactic parsing, although the user can annotate any linguistic feature he is interested in, from semantics to pragmatics, stylistic and error annotation.

In our case, corpus mark-up has already been provided by the creators of the corpora. What we are dealing with in this study is the annotation of modality, assigning a series of specific and manually created XML tags that provide information

about each modal marker. The modality tagger described in Chapter 5 follows the same process: from a raw text input, the program will automatically annotate the markers with XML. Table 10 shows an example of the modality annotation using XML in each corpora.

Table 10: Example of modality annotation (emphasis added for the example)

C-ORAL-ROM
<Utterance id="50" Type="enunciation"> **<m modtype="NEC" subtype="DEON" class="AUX" value="100%">**va a decir**</m>** mira pues hemos empezado a llamar pero como tú has sido el primero en llamar pues venga ya os **<m modtype="POSS" subtype="AMBG" class="AUX" value="50%">**podéis pasar**</m>** por aquí y **<m modtype="POSS" subtype="AMBG" elli="yes" class="AUX" value="50%"><v_elli type="poder"/>**pagar**</m>** </Utterance>
<Utterance id="1520" Type="enunciation"> **<w neg="yes">**no**</w>** me **<m modtype="NEC" subtype="DEON" class="AUX" value="0%" neg="yes">**deja comerlas**</m>** </Utterance>
<Utterance id="1708" Type="enunciation"> pero **<w neg="yes">**no**</w>** lo **<m modtype="NEC" subtype="DEON" class="mood_SUBJ" value="0%" neg="yes">**pongas**</m>** ahora </Utterance>
C-ORAL-JAPÓN
<UNIT id="75" speaker="TAK"> **それは <m modtype="POSS" subtype="EPIS" class="Adverb" value="50%"> たぶん </m> 下見じゃないですか？** </UNIT>
<UNIT id="277" speaker="MIZ"> **に** <utterance/> **な ん か 年 齢 登 録 <m modtype="NEC" subtype="DEON" class="AUX" value="100%"> しなきゃいけない </m> んでとかいってメール来て** <utterance/> **いや俺行かないけど** unkn **言ったら、**<utterance/> </UNIT>
<UNIT id="3376" speaker="YUK"> **何も <m modtype="NEC" subtype="DEON" class="mood_POT" neg="yes" value="50%"> とがめられない </m> から、**<utterance/> </UNIT>

3.3.2 Tagset used for the annotation

A tagset is a collection of tags and attributes of these tags that may or may not be obligatory to include. When a modal marker is encountered, it is annotated with an 'm' tag, followed by a series of attributes that classify and provide further information of the marker, all of them summarised below in Table 11. Five of these attributes are obligatory: main modality type ('modtype') which can be either 'NEC' (necessity) or 'POSS' (possibility); secondary modality type ('subtype') which can be 'DEON' (deontic), 'EPIS' (epistemic) or 'AMBG' (ambiguous); class of the modal marker ('class'), an auxiliary, an adverb, or an adjective; negative value, 'yes/no' depending on whether the marker is negated or not, and probability value ('value').

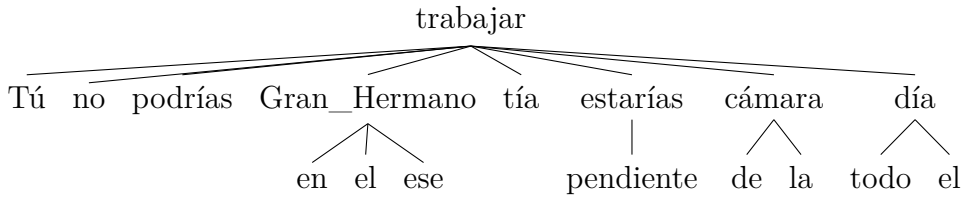
The case of negation presents two issues: first, as we have seen in Section 2.2.1, it may change the type of modality represented by the marker. Second, whereas in Japanese the negation is signalled by a suffix, in Spanish it is situated in a independent morpheme. Therefore, this morpheme will also be annotated with the 'word' tag ('w'). The modal marker that suffers the negation will include an attribute signalling it (neg="yes"). If the modality type is changed, the annotation will reflect it, tagging the marker with the new type. For example, the periphrastic construction *poder* + V ('may', 'can') belongs to the POSS modality as it denotes a possibility. However, when it is negated, it becomes an impossibility, that is, a NEC modality. Example 58 shows the annotation of a modal marker, including the negative element in an utterance of the corpus[4]. The marker has changed from a POS to a NEC modality.

[4]Utterance id: 22 of the corpus. Speaker: ROS

(58) a. *Tú no pod-rías trabaj-ar en el Gran Hermano*
you NEG can-COND.MODAUXNEG work-INF at the Big Brother
ése tía porque todo el día est-arías pendiente de la cámara
that mate because every the day be-COND waiting for the camera

'You couldn't work at Big Brother mate because you would be looking after the camera all day'

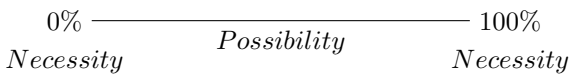
b.



c. tú **<w neg="yes">no</w> <m modtype="NEC" subtype="AMBG" class="AUX" value="0%" neg="yes">podrías trabajar </m>** en el Gran Hermano ése tía porque todo el día estarías pendiente de la cámara</Utterance>

The 'value' attribute is a percentage given to the marker indicating the probability it expresses approximately. This will be useful when comparing inter-linguistically each marker. If we situate modality values on a cline, necessity values would be on the extremes, indicating either 0% or 100%. Everything situated between these values will be considered a possibility marker, with a probability set at three possible values of 30%, 50% and 70% (see Figure 10). The idea has been influenced by the study made by Kawazoe et al. (2010) on aspect and certainty markers in Japanese.

Figure 10: Modal marker's probability percentage

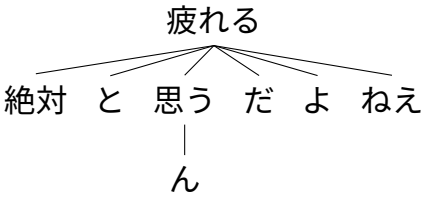


Example 59c. shows an example of annotating unit 59a. of the corpus[5]:

(59) a. 絶対 疲れ-る と 思-う ん だ よ
zettai *tsukare-ru* *to* *om-ō* *n* *da* *yo*
definitely.MODADV to get tired-PLN QUOT think-PLN EXPL COP EMPH
ねえ
nē
INT

'(I) think (he/she) will certainly get tired, right?'

b.



c. **<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">絶対</m>** ね疲れると思うんだよねえ、

Alternatively, there are three attributes that are optional: ellipsis of an element of the marker, separation of the elements that form the marker and *errors* made by the speaker. The first one is concerned with the possibility that part of the modal marker is missing due to an ellipsis made by the speaker voluntarily, or an ellipsis forced by an interruption or overlap by the hearer. This is highly probable in Spanish since periphrastic constructions made up by two or more independent words. One of the most frequent cases of ellipsis in Spanish periphrases can be found again in *poder* + V, for example in the sentence:

[5]UNIT id: 2088 of the corpus. Speaker: MIZ

(60) *¿Pued-o ir al cine? Claro que pued-o*
can-PRES.MODAUX go.INF to the cinema course CONJ can-PRES.MODAUX

'Can I go to the cinema? Of course I can (go)'

Although the main verb is missing in the second periphrasis, it is still a periphrasis and, therefore, annotated. In these cases the attribute 'ellipsis' is included (elli="yes") in the modality tag, and also a secondary tag 'v' is added signalling the lemma of the verb that has been omitted. For example, sentence 60 would be tagged as:

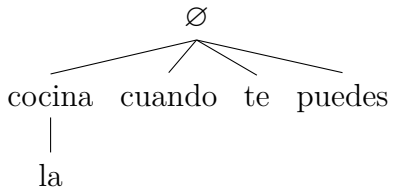
(60b.) ¿<m modtype="POS" subtype="AMBG" class="AUX" value="50%"> Puedo ir</m> al cine? Claro que **<m modtype="POS" subtype="AMBG" class="AUX" value="50%" elli="yes"><v_elli type="poder"/>puedo </m>**.

Example 61 shows another example of the annotation of an interrupted marker in a sentence from C-ORAL-ROM[6]. In this case the main verb does not appear in the corpus and it is labelled as 'inf':

(61) a. *La cocina cuando te pued-es ∅*
the kitchen when you.CLITIC can-PRES.MODAUXELLI INFELLI

'The kitchen when you can'

b.



c. la cocina cuando te **<m modtype="POSS" subtype="AMBG" class="AUX" value="50%" elli="yes">puedes<v_elli type ="inf"/> </m>**

[6]Utterance id: 2573 of the corpus. Speaker: RIC

Also, there is the case of coordination, which is partially treated as an ellipsis. The auxiliary verb could be followed by two main verbs coordinated (e.g. *tienes que cantar y bailar*, 'you have to sing and dance') and vice-versa (e.g. *puedes y debes bailar*, 'you can and must dance') or both at the same time. In these cases we would have two modal markers, one of them annotated as partially omitted. For example, in 'you have to sing and dance', there are two markers: 'have to sing" and 'have to dance'. The second has its auxiliary omitted, along with the connective preposition. The annotation process will be the same as the previous 60 and 61.

The second optional tag refers to those auxiliaries that may appear separated from the main verb in the same sentence, due to the inclusion of a clarification o hesitation elements. This happens only with auxiliaries and main verbs; the former will be tagged with the attribute 'ref' and the latter with 'id', as shown in the next example[7]:

(62) a. 立つ-ということ は あのう でき-ない

tatsu-toiukoto *wa* *anō* *deki-nai*

stand up.PLN-NMZ.REF_1 NOM well can-NEG.MODAUXNEGID_1

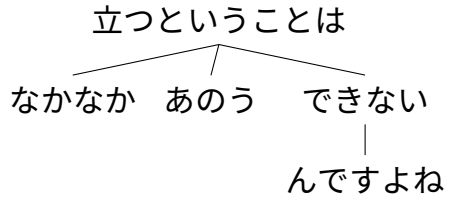
ん です よ ね

n *desu*COP *yo* *ne*

EXPL EMPH INT

'(you) cannot quite, well, such thing as standing up, right?'

b.



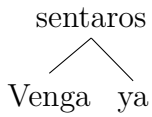
c. **<m class="AUX" id="1" modtype="NEC" neg="yes" subtype="DEON" value="0%">立つということは</m>** なかなか **あのう <m class="AUX" ref="1" modtype="NEC" neg="yes" subtype="DEON" value="0%">できないん</m>** ですよね

[7]Utterance id: 13614 of the corpus. Speaker: HOS

Finally, we may find in the spoken register *errors* made by the speakers. It is outside the scope of this study the discussion about whether or not these cases are in fact 'mistakes' or 'errors' or simply instances of possible cases of language change (where the label 'error' would bare a negative connotation). They are considered 'errors' only because they do not follow the original rule, with no negative implication intended. The most common errors take place in Spanish between the constructions *Deber* + V ('must', deontic) and *Deber de* + V ('must', epistemic) and using an infinitive verb as an imperative. In these cases, the construction is left untouched and labelled as a modal marker, but with the attribute 'error' inside the tag. Sentence 63 taken from the corpus[8] shows an example. The verb *sentaros* ('sit down') is an infinitive form of *sentarse* but it is used as the imperative *sentaos*:

(63) a. *Venga* *sentaros* *ya*
come on sit down.MODIMP_ERR already

'Come on sit down already'

b. 

c. Venga **<m modtype="NEC" subtype="DEON" class="mood_IMP" value="100%" error="yes">sentaros</m>ya**

Aside from the modality tags ('m'), there are non-modal elements of the sentence that also are annotated with 'w'. One is the negative element, as explained a few lines above; the other is a discourse marker. In Spanish there are discourse markers that may be erroneously annotated by the tagger as modal markers, since they have the same structure as a periphrastic construction. One example of this is *vamos a ver* ('let's see...') which is formed by an inflected form of the verb *ir* ('to

[8]Utterance id: 1645 of the corpus

go') + connective 'a' + V in infinitive form, just like the modal periphrasis *ir a* + V. These cases have been annotated with the word tag 'w" followed by the MD type attribute (type="MD").

Table 11 summarises the XML tags and attributes used for the study. We have seen in this section how XML works and what is the tagset used in the annotation and by the automatic tagger in this study. The next section (3.4) will finish describing the tools used for this study, the automatic POS taggers for each language, and programming language, before listing and discussing each modal marker and their XML tag in Section 3.5, which will conclude the chapter.

Table 11: XML tags used for the annotation

Tag/Element	**Attributes**	
	Name	**Possible values**
m	Modtype [Obligatory]	- NEC (Necessity) - POSS (Possibility)
	Subtype [Obligatory]	- EPIS (Epistemic) - DEON (Deontic) - AMBG (Ambiguous)
	Class [Obligatory]	- AUX (Auxiliary: Periphrases/Suffixes) - Adverb - Adjecive - mood_IMP (Imperative mood) - mood_SUBJ (Subjunctive mood)
	Value [Obligatory]	- 0%, 30%, 50%, 70%, 100%
	Neg [Obligatory]	- Yes/No
	Elli [Optional]	- Yes
	ID/ref [Optional]	- 1
	Error [Optional]	- Yes
w	Type [Obligatory]	- Neg (Negative element) - MD (Discourse marker)
v_elli	Type [Obligatory]	- Open value (Lemma of omitted verb) - Inf (Infinitive verb omitted)

3.4 Tools

3.4.1 POS taggers

3.4.1.1 Grampal, a tagger for Spanish

After the annotation in the corpora is made, the modality tagger will try to reproduce the annotation automatically in new raw text through a series of hand-created rules based on the information learned from the theory and corpus findings. These rules will benefit from two POS taggers, one for each language. Not only they will improve the efficiency and accuracy of the program, but also make the process simpler, especially for lemmatising, finding elements in the sentence such as the subjunctive or imperative moods, etc. If the tools are already created, it would be unwise not to use them and build new things upon them. Also, nowadays the usage of POS tagged text for NLP activities is, and should be, considered the first stage of any form of annotation (Leech & Smith, 1999).

A part-of-speech or POS tagger is a software that automatically assigns a word its word-class information, as well as a morphosyntactic analysis such as tense, mood, aspect, number, person, etc. Some taggers, such as Juman, can provide us with additional features such as the Latin reading of Japanese characters (romaji) or even the semantic domain of the word. A POS tagger would automatically tell us if *pencil* is a singular masculine noun or *perhaps* is an adverb. It is normally formed by three modules (Voutilainen, 1999, p. 6):

1. A *tokeniser* that separates the words of a text. Some POS taggers may recognise multiword expressions, idioms that work as single words.
2. A *lookup* module that assigns the morphosyntactic analysis. It is comprised by a *lexicon*, collecting stems and affixes, and a *guessing* module that analyses words that do not appear in the lexicon.
3. A *disambiguation* module that predicts and chooses the correct analysis in the case of multiple possible ones.

In the Spanish scenario, the automatic annotation of modality will mostly take advantage of the detection by the POS tagger of open-class elements, like verbs and their subjunctive and imperative moods. If we did not have a POS tagger, we would have to create manually rules for detecting these elements. We have selected for this purpose the Grampal[9] tagger for Spanish (Moreno & Goñi, 1995). The most significant characteristic of this program is its ability to detect proper nouns and some multiword expressions. Also, it has been especially trained for the spoken discourse, and can detect discourse markers (Moreno & Guirao, 2003, 2006). Example 64 shows an example of the output made by this tagger (Input sentence: *A lo mejor como en casa mañana.* 'Tomorrow I will probably eat at home'). Grampal separates the form, the lemma, the wordclass tag, and the morphosyntactic information:

(64) A lo mejor/A LO MEJOR/ADV como/COMER/V/sing,1,pres_ind en/EN/PREP casa/CASA/N/fem,sing mañana/MAÑANA/N/sing

Another important feature of Grampal is its small amount of tags used. The objective was to maintain a reduced tagset of 18 main tags for a simpler annotation, followed by subtags indicating person, number, tense and mood.

1. N Noun
2. NPR Proper Noun
3. ADJ Adjective
4. V Verb
5. AUX Auxiliary
6. P Pronoun
7. REL Relative
8. PINT Interrogative Pronoun
9. ART Article
10. POSS Possessive
11. DEM Demonstrative

[9] http://cartago.lllf.uam.es/grampal/grampal.cgi

12. Q Quantifier
13. PREP Preposition
14. ADV Adverb
15. C Conjunction
16. INTJ Interjection
17. MD Discourse Marker
18. UNKN Unknown word

The most useful tags for the automatic implementation will be Vsubj and Vimper for detecting subjunctive and imperative moods, as well as MD for separating discourse markers from periphrastic constructions.

3.4.1.2 Juman, a tagger for Japanese

For Japanese, a POS tagger is especially useful for the tokenisation of words. It is well known that the principal obstacle in Japanese NLP is the lack of white spaces between words, making the automatic handling of data and annotation extremely difficult. To tackle this problem and nearly any Japanese NLP exercise, it is necessary to use a Japanese tagger that separates for us the words so we can perform the analysis. This process, however, is not perfect. The very definition of *word* in Japanese presents its problems and there is not a united view on the matter. For example, as we saw in Section 2.3.1, the auxiliary かもしれない (*kamoshirenai* can be considered a single entity, or divided into three smaller auxiliaries, かも (*kamo*), しれ (*shire*) and ない. This leads to Japanese POS taggers performing different segmentations and tokenisations of a text (Asahara et al., 2002). Also, a consequence of this is the typical problem of *oversegmenting* a text (Hisamitsu & Nitta, 1996), overdividing words resulting in erroneous separations, especially in compound words. For example, word 鼓室 'tympanic cavity' may be divided into 鼓 'hand drum' and 室 'room', two different words that lose the medical meaning (Herrero, 2013a). Although the spoken discourse will not have a high amount of specialised compounds, we may encounter this problem.

There are three main Japanese POS taggers: ChaSen(Matsumoto et al., 2002)[10], Mecab[11] and Juman[12]. After several tests on written discourse, we observed that all of them separate words in similar ways. However we concluded that Juman was the best one for our purposes because it provides the widest amount of information for each word (Herrero et al., 2014). For this study, Juman (Matsumoto et al., 1997) specifically becomes a useful tool since it provides the mood and the type of stem of the verb, which will help the detection of modal auxiliaries by the automatic tagger (See Chapter 5). Example 65 shows an example of the output made by Juman (input sentence: 生の魚が食べたい 'I want to eat raw fish'.). Juman provides the form, hiragana reading, lemma, tag, subtag, and information such as the type of inflection of the verb (in this case 連用形, 'adverbial' or 'continuative form') and the semantic domain of the word (料理, 'food', for 'fish'):

(65) 生せい生名詞 6 普通名詞 1 * 0 * 0 '代表表記: 生/せい漢字読み: 音カテゴリ: 抽象物"
@ 生なま生名詞 6 普通名詞 1 * 0 * 0 '代表表記: 生/なま漢字読み: 訓カテゴリ: 抽象物"
ののの助詞 9 接続助詞 3 * 0 * 0 NIL
魚ぎょ魚名詞 6 普通名詞 1 * 0 * 0 '代表表記: 魚/ぎょ漢字読み: 音カテゴリ: 動物ドメイン: 料理・食事"
@ 魚さかな魚名詞 6 普通名詞 1 * 0 * 0 '代表表記: 魚/さかな漢字読み: 訓カテゴリ: 動物ドメイン: 料理・食事"
ががが助詞 9 格助詞 1 * 0 * 0 NIL
食べたべ食べる動詞 2 * 0 母音動詞 1 基本連用形 8 '代表表記: 食べる/たべるドメイン: 料理・食事"
たいたいたい接尾辞 14 形容詞性述語接尾辞 5 イ形容詞アウオ段 18 基本形 2 '代表表記: たい/たい"

[10]http://chasen-legacy.sourceforge.jp/
[11]http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[12]http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

Juman's selection of tags has been inspired from Masuoka and Takubo's grammar (1992). As Grampal, it contains a short tagset of 14 different tags, which are followed by a different array of subtags and sometimes *subsubtags*, as seen in Example 65. The automatic modality tagger will benefit from them, especially for the detection of the main and auxiliary verbs (as well as the type of inflection), adjectives and adverbs.

1. 動詞 Verb
2. 助動詞 Auxiliary verb
3. 形容詞 *Pure* adjective (-i adjective)
4. 連体詞 Adjective (-na adjective)
5. 判定詞 Special (-no) adjective
6. 名詞 Noun
7. 指示詞 Demonstrative
8. 副詞 Adverb
9. 助詞 Particle
10. 接続詞 Conjunction
11. 感動詞 Interjection
12. 接頭辞 Prefix
13. 特殊 Special (symbol)
14. 未定義語 Unknown word

3.4.2 Python Language and Prism

Finally, for the manipulation of data, text processing, statistical calculations and representation, and overall design of the modality tagger, we used the *Python Programming Language* (Python Software Foundation, 2016). The version selected was Python 3 due to its ability for handling Unicode characters. Additionally, the following libraries where used for the main tasks:

- XML parsing: *LXML* (lxml Project, 2016) and *Beautiful Soup* (Richardson, 2016)
- Database handling: *Pandas* (McKinney, 2010)
- Plotting: *Matplotlib* (Hunter, 2007)

Lastly, Graphpad's *Prism* software version 5.0 for MacOS[13] was used for plotting of some of the graphs and calculation of normality and t tests. This ends the section dedicated to the main tools used for the study: XML language, a POS tagger for each language, and *Python* language, including several specialised libraries. The next section will list and describe each modal marker considered for this study as well as the tags and attributes assigned to them.

[13]GraphPad Software, La Jolla California USA, www.graphpad.com

3.5 Modal markers

Each and every modal marker encountered has been tagged in both corpora with the same set of XML tags in both languages: type of modality (necessity or possibility), subtype of modality (deontic, epistemic or ambiguous), class of marker (periphrasis, suffix, adverb, adjective or mood), if it is negated or not, and degree of probability.

The Spanish markers were compiled mainly from resources from Spanish grammars and dictionaries lite Kovacci (1999); Gómez Torrego (1999); RAE (2001); RAE (2009) as well as other works such as (Gómez Manzano, 1991; Kaul de Marlangeon, 2002; Cornillie, 2010). Japanese do not have an *official* grammar per-se, but modal markers were compiled from other works with quantitative, corpus linguistics studies in mind such as Larm (2006); Narrog (2009a); Kawazoe et al. (2010), with special reference to Matsuoka (1981), and the dictionary of Japanese Function Expressions (Matsuyoshi et al., 2007). Each marker was classified under necessity or possibility, and given the appropriate information regarding the modality subtype, the grammatical class, etc.

This section will be divided into three main parts depending on the nature of each marker: auxiliaries, adverbs and adjectives.

3.5.1 Auxiliaries

3.5.1.1 Spanish auxiliaries

As explained in Chapter 2, Spanish auxiliary verbs can create periphrastic constructions formed by the auxiliary in finite form, which contains subject agreement, tense, aspect and mood information, followed by the main verb in non-finite form, sometimes joined by a connective. There are many types of periphrases in Spanish, but only the modal ones will be studied in this work. These constructions are limited to 7 auxiliaries, with an open-class main verb that has to be in infinitive form. We will stop and look at each one in the next pages:

1. *Poder* + V ('can', 'may')
2. *Deber* + V ('must', 'have to')
3. *Deber de* + V ('must', 'have to')
4. *Haber de* + V ('must', 'have to')
5. *Haber que* + V ('must', 'have to')
6. *Tener que* + V ('must', 'have to')
7. *Ir a* + V ('will', 'going to')

Poder + V

The periphrasis formed by the auxiliary *poder* is the most frequent, as seen below, and also the one that contains more different meanings, becoming the most ambiguous one. It represents a possibility, but it can either be of an event happening (epistemic reading), a permission or an ability (deontic readings). This construction can naturally omit the main, non-finite verb in standard Spanish, both written and spoken. Finally, regarding the issue of negation, it can be negated in the auxiliary but also the main verb position. In other words, negation can affect the modal auxiliary, but also the proposition. Both sentences 66a. 66b. are grammatical:

(66) a. *No pued-o ir mañana a clase*
NEG can-PRES.MODAUXNEG go.INF tomorrow to class

'I cannot go to class tomorrow'

b. *Pued-es no ven-ir si no quier-es*
can-PRES.MODAUXID_1 NEG come-INFREF_1 if NEG want-PRES

'You don't have to come if you don't want to'

In 66a., the periphrasis *Poder* + V involves an impossibility. The negation affects modality and transforms the possibility into a necessity, following the rule $\Box p \iff \neg\Diamond p$ (seen in Chapter 2). In 66b., *Poder* + V remains a possibility. The negation only affects the proposition 'to come'. That is $\Diamond p \iff \Diamond\neg p$. Table 12 resumes the characteristics of this construction:

Table 12: Information for Poder + V

Poder + V	
Issue	**Result**
English Equivalents	Can, could, may, might
Modality Type	Possibility
Modality Subtype	Ambiguous
Negation Change	Yes
Negation of Auxiliary	Possible
Negation of Proposition	Possible
Probability Percentage	50%, 0% if negated
Obligatory Tag	<m modtype="NEC" subtype="AMBG" class="AUX" value="50%">

Deber (de) + V

This construction has a single obligatory meaning and therefore represents a necessity modality. It also presents no ambiguity and is classified as a Deontic marker. However, there are two representative features regarding this construction. First, in terms of negation, the auxiliary can receive a negative element and also the proposition, like the previous construction (Giammatteo & Marcovecchio, 2010). That is, sentences 67a. and 67b. are grammatical. However, on the contrary to what happens with *Poder* + V, even though modality is negated, the necessity is maintained. The rule $\Diamond p \iff \neg\Box p$ does not apply.

(67) a. *No deb-es venir mañana a clase*
not.NEG must-PRES.MODAUX go.INF tomorrow to class

'You must not come to class tomorrow'

b. *Deb-es no venir mañana a clase*
must-PRES.MODAUXID_1 NEG go.INFREF_1 tomorrow to class

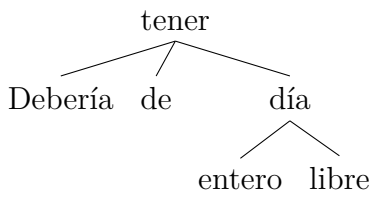
'You must not come to class tomorrow'

Secondly, this periphrasis presents confusions among speakers with another one, *Deber de* + V, as seen in Section 3.3.2. Although both of them are very similar, the latter includes preposition *de* as a connective between the auxiliary and the main verb, and the meaning is different. Both can be translated to English as 'must', but whereas *Deber* + V has the modal meaning of deontic necessity, *Deber de* + V involves an epistemic modality, that is, a hypothesis made by the speaker of a specific event. Both constructions are seldom used interchangeably in spontaneous spoken Spanish as some studies have pointed out before (Gómez Manzano, 1991; Herrero, 2014) and will be confirmed in this study. Nevertheless, the confusion also appears to be present already in some classic texts and in the literary context (RAE, 2009, p. 239), which leads us to believe that in the future both constructions will blend completely. The next Example 68 shows this issue in a sentence from the corpus[14].

[14]UNIT id: 5935 of the corpus. Speaker: MAY

(68) a. *Deb-ería* *de* *tener* *un* *día*
must-COND.MODAUX_ERR of.CONN_ERR have.INF_ERR one day
entero *libre*
whole free

'I should have a whole free day'

b. 

c. **<m modtype="NEC" subtype="DEON" class="AUX" value="100%" error="yes">debería de tener</m>** un día entero libre

Even though it can be considered an epistemic construction, its overlapping with modal *Deber* + V is clear. In the annotation process we will treat them separately and mark the ones that have been used erroneously to observe this overlapping (Tables 13 and 14).

Table 13: Information for Deber + V

Deber + V	
Issue	**Result**
English Equivalents	Must, should
Modality Type	Necessity
Modality Subtype	Deontic
Negation Change	No
Negation of Auxiliary	Possible
Negation of Proposition	Possible
Probability Percentage	100%
Obligatory Tag	<m modtype="NEC" subtype="DEON" class="AUX" value="100%">

Table 14: Information for Deber de + V

Deber de + V	
Issue	**Result**
English Equivalents	Must, should
Modality Type	Necessity
Modality Subtype	Epistemic
Negation Change	No
Negation of Auxiliary	Possible
Negation of Proposition	Not possible
Probability Percentage	100%
Obligatory Tag	<m modtype="NEC" subtype="EPIS" class="AUX" value= "100%">

Tener que + V

This construction (15), which contains the conjunction *que* linking the auxiliary and the main verb, has a very similar necessity meaning to the previous *Deber* + V, reaching a semantic neutralisation between them that allows the speaker to use them indistinctly (Gómez Manzano, 1991, p. 155). There are, however, some differences. First, this marker can have epistemic reading, becoming, like *Poder* + V, ambiguous. The epistemic reading expresses a 'certain inference' or 'conclusion', and the deontic an obligation or an inevitable necessity (RAE, 2009). Secondly, when pre-negated, the modality changes to a possibility ('there is no need to...').

Table 15: Information for Tener que + V

Tener que + V	
Issue	**Result**
English Equivalents	Must, should
Modality Type	Necessity
Modality Subtype	Ambiguous
Negation Change	Yes
Negation of Auxiliary	Possible
Negation of Proposition	Not possible
Probability Percentage	100%, 50% if negated
Obligatory Tag	<m modtype="NEC" subtype="AMBG" class="AUX" value="100%">

Haber que + V

Haber que + V is very similar to the previous one: it denotes a necessity and contains a link in-between the auxiliary and main verbs, and it has a deontic reading as it imposes the necessity to the receiver of the message. The peculiarity of this construction is its limited use to the third person, acquiring a general and impersonal necessity (Gómez Manzano, 1991, p. 166). Also, when the main verb is *ver* ('to see') it may function as a lexicalised discourse marker construction that has lost its modal meaning (*hay que ver*, exclamatory expression e.g. 'fancy that', 'good heavens'). Finally, as with *Tener que*, the negation changes the type of modality (Table 16).

Table 16: Information for Haber que + V

Haber que + V	
Issue	**Result**
English Equivalents	Must, should
Modality Type	Necessity
Modality Subtype	Deontic
Negation Change	Yes
Negation of Auxiliary	Possible
Negation of Proposition	Not possible
Probability Percentage	100%, 50% if negated
Obligatory Tag	<m modtype="NEC" subtype="DEON" class="AUX" value="100%">

Haber de + V

This expression continues signalling necessity as the previous periphrases (Table 17). It is very similar to *Tener que*, as it has epistemic and deontic meanings, but it does not change with negation, similar to *Deber* + V. It is the oldest construction (Gómez Manzano, 1991, p. 169), and it is limited almost exclusively to the written discourse.

Table 17: Information for Haber de + V

Haber de + V	
Issue	**Result**
English Equivalents	Must, should
Modality Type	Necessity
Modality Subtype	Ambiguous
Negation Change	No
Negation of Auxiliary	Possible
Negation of Proposition	Not possible
Probability Percentage	100%
Obligatory Tag	<m modtype="NEC" subtype="AMBG" class="AUX" value="100%">

Ir a + V

The construction *Ir a* + V (Table 18) is a problematic one, as its modal value is questionable, depending on the grammarians and linguists that study it. The main issue is its future meaning that overlaps with the modal meaning of 'will' or 'intention'. It can be translated to English as 'will' but also 'going to'. Some grammars signal it exclusively as a temporal construction (RAE, 2009, p. 541), and others do not talk about its future value or even classify it as a periphrastic construction (Gómez Manzano, 1991, p. 109). Here it will be considered as a modal construction that includes future meanings. In general, both modal and temporal senses are enclosed and cannot be separated, although in some cases, such as with

atmospheric verbs (*va a llover* 'it is going to rain'), they can only represent future or evidential cases. An intention or will involves a state of affairs signalled as necessarily true by the speaker but has yet been realised, acquiring the future sense. In terms of negation, it will maintain necessity when it is negated in a previous position. Also, in a similar way to *Haber que*, it may appear as a lexicalised modal discourse when the main verb is *ver* ('to see') and the auxiliary on the first person plural: *vamos a ver* ('let's see...').

Table 18: Information for Ir a + V

Ir a + V	
Issue	**Result**
English Equivalents	Will, going to
Modality Type	Necessity
Modality Subtype	Deontic
Negation Change	No
Negation of Auxiliary	Possible
Negation of Proposition	Not possible
Probability Percentage	100%, 0% if negated
Obligatory Tag	<m modtype="NEC" subtype="DEON" class="AUX" value="100%">

To summarise, Spanish can signal modality through seven different periphrastic constructions. The most characteristic ones are *Poder* + V and *Deber* + V as the ones who are compatible with proposition negation. Regarding possibility and necessity, the auxiliary *Poder* is the only one that encodes possibility. In terms of epistemic and deontic readings, there is ambiguity in three constructions, *Poder*, *Tener que* and *Haber de*. The other three have deontic meanings. Also, there is not a construction that will only encode epistemic modality. Finally, all of them can receive previous negation, three constructions change to necessity or possibility when negated (*poder*, *tener que*, *haber que*). The rest (*beber*, *haber de*, *ir a*) remain unchanged. The next section will reproduce the analysis with Japanese modal auxiliaries.

3.5.1.2 Japanese auxiliaries

Chapter 2 showed how Japanese auxiliaries are attached to and modify the main verb in a very similar way to Spanish periphrastic constructions, although they have been grammaticalised into bound morphemes. They can be divided into periphrastic constructions and suffixes according to their degree of grammaticalisation, but it seemed reasonable to simplify and group them into a single notion of auxiliaries, for the sake of simplicity, as some authors have previously done.

The first noticeable difference from their Spanish counterparts is the amount of modal auxiliaries available in Japanese, up to 23. This will lead to more specialised meanings, and in terms of the more general classification between necessity, possibility, epistemic and deontic, the overlapping is minimum. The second difference is the variation according to register, gender and writing. There are many morphological and lexical variations in Japanese according to the register (depending on the social status, age, proximity, etc. of the hearer), gender of the speaker, and type of discourse (spoken or written). This also applies to modal auxiliaries, and the study must take into account these different possibilities for each marker. The same goes with the type of writing, as some of them can be written in kanji or hiragana, depending on the choice of the writer, or the transcriber of the spoken corpus.

From the 23 Japanese modal auxiliaries, which spread up to nearly 60 different forms if we take into account formality and discourse variations, only 7 encode possibility. Between the 16 remaining expressions dedicated to necessity, 6 of them express epistemic notions. That is, there are at least 10 grammaticalised expressions that express deontic necessity, i.e. obligations, expressions that the speaker uses to impose a state of affairs upon the speaker or him/herself. Although some of them are used interchangeably, there are small theoretical differences and restrictions according to the social situation that select one or another, as we saw in the Introduction. This variability also leads to a lesser amount of ambiguity, as we could only find overlapping between deontic and epistemic readings in one expression. We will begin our analysis with the necessity expressions, before addressing the possibility ones. The information has been extracted from different Japanese

studies (Matsuoka, 1981; Makino & Tsutsui, 1994, 1995, 2008; Matsuyoshi et al., 2007; Narrog, 2009a; Kawazoe et al., 2010) as well as observations from the corpus.

To begin with, there are three auxiliaries that denote a strong obligation: なければならない (*nakerebanaranai*), ざるを得ない (*zaruwoenai*), やむを得ない (*zaruwoenai*), しかない (*shikanai*).

V + なければならない (*nakerebanaranai*)

This construction involves a strong obligation towards the receiver of the message made by the speaker, or, in other words, a deontic modality (Table 19). There are some cases where it can be used to signal epistemic modality, and some authors believe these are premature signals indicating it may become an epistemic marker in the future (Narrog, 2012). Considering the semantic maps of modality (see Figure 3 of Section 2.2.2), it would not be surprising. Nevertheless, for the time being, it will be considered exclusively as deontic. It is also one of the most used modal markers in Japanese and can be rephrased into 30 different expressions including formal and informal ones. Although it has become a single expression with a specific meaning, its basic form is built from a combination of auxiliary verbs and adjectives, meaning literary 'the not possibility of not V'. The example 69 shows this with the construction 'Must go':

(69) 行-か な-ければ なら-ない
ika *na-kereba* *nara-nai*
go-IRR NEG-if become-NEG

'(Subject) must go'

The expression itself is a negation of a possibility, becoming a necessity. The auxiliaries can change into more formal/informal alternatives, but combine together to form the same meaning, if only mitigating in some cases the degree of obligation. A negation of this marker is not possible, and neither the negation of the proposition.

Table 19: Information for V + なければならない

V + なければならない	
Issue	**Result**
English Equivalents	Must, have to
Modality Type	Necessity
Modality Subtype	Deontic
Negation Change	No
Negation of Auxiliary	Not possible
Negation of Proposition	Not possible
Probability Percentage	100%
Obligatory Tag	<m modtype="NEC" subtype="DEON" class="AUX" value="100%">

V + ざるを得ない (*zaruwoenai*) and V + しかない (*shikanai*)

These auxiliaries also impose the state of affairs to the receiver of the message, although they involve a sense of inevitability and not providing an alternative choice, not present in the previous one. In terms of negation, they are also composed by negative elements. The first one (Table 20) is formed by a double negation with ざる (*zaru* 'not') and 得ない (*enai* negative of 'to acquire' or 'be able to'), literally meaning 'the not V is not possible', or the impossibility of p not becoming true that becomes the necessity of V. It can be formalised as ¬◇¬p which equals to □p. The second (Table 21) is a negation of しか (*shika*) 'only, but', meaning as a whole 'nothing but'.

This means they do not accept additional negative elements to the construction making it very similar, grammatically, to the previous *nakerebanaranai*. Also, there is an alternative realisation of this expression with the negative やむ (*yamu*, やむを得ない).

Table 20: Information for V + ざるを得ない

V + ざるを得ない	
Issue	**Result**
English Equivalents	Must, have to
Modality Type	Necessity
Modality Subtype	Deontic
Negation Change	No
Negation of Auxiliary	Not possible
Negation of Proposition	Not possible
Probability Percentage	100%
Obligatory Tag	<m modtype="NEC" subtype="DEON" class="AUX" value="100%">

Table 21: Information for V + しかない

V + しかない	
Issue	**Result**
English Equivalents	Must, have to
Modality Type	Necessity
Modality Subtype	Deontic
Negation Change	No
Negation of Auxiliary	Not possible
Negation of Proposition	Not possible
Probability Percentage	100%
Obligatory Tag	<m modtype="NEC" subtype="DEON" class="AUX" value="100%">

V + 訳にはいかない (*wakenihaikanai*)

訳にはいかない, and its many alternatives –訳にはいけない (*wakenihaikenai*), てはいか (け) ない (*tehaika(ke)nai*), てはならない (*tehanarani*), てはいられない (*tehairarenai*)– also involve an imposition on the receiver of the message, but in the form of an impossibility, similar to English 'cannot' (Makino & Tsutsui, 1994). The items that compose this expression, V + 訳には (*wakeni* 'the circumstance of') + いかない (*ikanai* 'not reaching'), once again form a negation *per se* literally meaning 'the impossibility of V'. This construction is compatible, however, with a negation of p. In other words, it can be attached to a main verb in its negative form (Table 22).

Table 22: Information for V + 訳にはいかない

V + 訳にはいかない	
Issue	**Result**
English Equivalents	Cannot, must not
Modality Type	Necessity
Modality Subtype	Deontic
Negation Change	No
Negation of Auxiliary	Not possible
Negation of Proposition	Possible
Probability Percentage	0%
Obligatory Tag	<m modtype="NEC" subtype="DEON" class="AUX" value="0%">

V + に忍びない (*nishinobinai*)

This construction refers to an internal obligation of the speaker not being able to achieve a given state of affairs (Table 23). It could be translated as 'could not bring oneself into something', roughly equivalent to English 'cannot' or 'could not'. It is again an auxiliary in negative form, and cannot accept additional negation of any kind.

Table 23: Information for V + に忍びない

V + に忍びない	
Issue	**Result**
English Equivalents	Cannot, could not
Modality Type	Necessity
Modality Subtype	Deontic
Negation Change	No
Negation of Auxiliary	Not possible
Negation of Proposition	Not possible
Probability Percentage	0%
Obligatory Tag	<m modtype="NEC" subtype="DEON" class="AUX" value="0%">

V + べき (*beki*), **V + た方がいい** (*tahōgaii*), **V + たらいい** (*taraii*), **V + ればいい** (*rebaii*)

The following four markers also express necessity and deontic values, but they are used for giving recommendations or mitigated obligations. The state of affairs is marked as appropriate morally or socially by the speaker and involves warnings, desires or wishes if the SOA is uncontrollable. The first one, べき (*beki*) (Table 24) is an auxiliary that means 'should' or 'ought to', involving a 'duty' or 'obligation'. The rest (Tables 25, 26 and 27) are combinations of the conditional form of the verb followed by adjective 'good' (いい), literally 'it is good if V'. In summary, they are expressions related to what one is supposed to do in the society he/she lives in. The speaker imposes the SOA depending on what is socially or morally expected, contrasting with former *nakerebanaranai* which is an imposition made by the speaker on his/her own terms (Imithani, 2009, p. 60). There are some differences in the negation. Regarding negation of the proposition, it is possible in the three first べき (*beki*), た方がいい (*tahōgaii*) and たらいい (*taraii*). These too accept negation of the auxiliary, but only in interrogative sentences such as in Example 70 to confirm the message with the receiver. The modal is not negated in these cases.

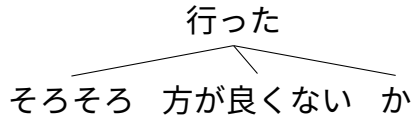
(70) a. そろそろ 行っ-た-方が良-くない か

sorosoro *it-ta-hōgayokunai* *ka*

soon go-PST-have-NEG.MODAUXNEG INT

'Shouldn't we go?'

b.



c. そろそろ **<m modtype="NEC" subtype="DEON" class="Aux" neg="no" value="100%">行ったほうがよくない** </m> か

Table 24: Information for V + べき

V + べき	
Issue	**Result**
English Equivalents	Should, ought to
Modality Type	Necessity
Modality Subtype	Deontic
Negation Change	No
Negation of Auxiliary	Not possible
Negation of Proposition	Possible
Probability Percentage	100%
Obligatory Tag	<m modtype="NEC" subtype="DEON" class="AUX" value="100%">

Table 25: Information for V + た方がいい

V + た方がいい	
Issue	**Result**
English Equivalents	Should, ought to
Modality Type	Necessity
Modality Subtype	Deontic
Negation Change	No
Negation of Auxiliary	Not possible
Negation of Proposition	Possible
Probability Percentage	100%
Obligatory Tag	<m modtype="NEC" subtype="DEON" class="AUX" value="100%">

Table 26: Information for V + たらいい

V + たらいい	
Issue	**Result**
English Equivalents	Should, ought to
Modality Type	Necessity
Modality Subtype	Deontic
Negation Change	No
Negation of Auxiliary	Not possible
Negation of Proposition	Possible
Probability Percentage	100%
Obligatory Tag	<m modtype="NEC" subtype="DEON" class="AUX" value="100%">

Table 27: Information for V + ればいい

V + ればいい	
Issue	**Result**
English Equivalents	Should, ought to
Modality Type	Necessity
Modality Subtype	Deontic
Negation Change	No
Negation of Auxiliary	Not possible
Negation of Proposition	Not possible
Probability Percentage	100%
Obligatory Tag	<m modtype="NEC" subtype="DEON" class="AUX" value="100%">

V + たい (*tai*), **V + てもらいたい** (*temoraitai*), **V + てほしい** (*tehoshii*), **V + てください** (*tekudasai*)

These five constructions involve desires of the speaker, either internal –**たい** (*tai*) and **てほしい** (*tehoshii*) 'I want/need V' (Tables 28 and 30); or external: desiring a state of affairs to be necessarily true through the receiver –**てもらいたい** (*temoraitai*) and **てください** (*tekudasai*) 'I want/need (receiver) to V' (Tables 29, 31). Hence, they are considered as necessity deontic modality markers. In terms of negation, all of them except **てください** (*tekudasai*) are compatible with negation in the auxiliary, but don't undergo modality change. Regarding negation of the proposition, all of them can be attached to a negative main verb except **たい** (*tai*).

We cannot find in Spanish modal markers equivalent to these. That is to say, the expressions involving desire, both internal and external, have yet to be grammaticalised in the same degree as their Japanese counterparts. The verb *querer* ('to want') appears to be in the middle of the process according to some linguists as it starts to function as an auxiliary joined to an infinitive verb in a pseudo-periphrastic construction like in *Quiero volar* 'I want to fly'. Some grammars have labelled these forms as 'semiauxiliaries' and even consider them a periphrastic construction (RAE,

2009) However, the process is not complete and it is still considered a lexical verb on its own for this study.

Table 28: Information for V + たい

V + たい	
Issue	**Result**
English Equivalents	Want/Need
Modality Type	Necessity
Modality Subtype	Deontic
Negation Change	No
Negation of Auxiliary	Possible
Negation of Proposition	Not possible
Probability Percentage	100%
Obligatory Tag	<m modtype="NEC" subtype="DEON" class="AUX" value="100%">

Table 29: Information for V + てもらいたい

V + てもらいたい	
Issue	**Result**
English Equivalents	Want/Need (the receiver to...)
Modality Type	Necessity
Modality Subtype	Deontic
Negation Change	No
Negation of Auxiliary	Possible
Negation of Proposition	Possible
Probability Percentage	100%
Obligatory Tag	<m modtype="NEC" subtype="DEON" class="AUX" value="100%">

Table 30: Information for V + てほしい

V + てほしい	
Issue	**Result**
English Equivalents	Want/Need (the speaker or receiver to...)
Modality Type	Necessity
Modality Subtype	Deontic
Negation Change	No
Negation of Auxiliary	Possible
Negation of Proposition	Possible
Probability Percentage	100%
Obligatory Tag	<m modtype="NEC" subtype="DEON" class="AUX" value="100%">

Table 31: Information for V + てください

V + てください	
Issue	**Result**
English Equivalents	Want/Need (the receiver to...)
Modality Type	Necessity
Modality Subtype	Deontic
Negation Change	No
Negation of Auxiliary	Not possible
Negation of Proposition	Possible
Probability Percentage	100%
Obligatory Tag	<m modtype="NEC" subtype="DEON" class="AUX" value="100%">

V + つもり (*tsumori*)

The modal marker つもり (*tsumori*), literary 'plan' or 'intention' is equivalent to Spanish *ir a* + V and English 'will' or 'going to' (Table 32). However, it only encodes the modal meaning. Whereas the Spanish and English expressions also involve a future and evidential sense, the Japanese is restricted to the deontic expression of intention of the speaker. The speaker expresses its intention to convert the state of affairs into something necessarily true. In terms of negation, the auxiliary can be negated as well as the proposition, but there is no change in modality.

Table 32: Information for V + つもり

V + つもり	
Issue	**Result**
English Equivalents	Will, going to
Modality Type	Necessity
Modality Subtype	Deontic
Negation Change	No
Negation of Auxiliary	Possible
Negation of Proposition	Possible
Probability Percentage	100%
Obligatory Tag	<m modtype="NEC" subtype="DEON" class="AUX" value="100%">

V + かねる (*kaneru*), V + はず (*hazu*), V + に違いない (*nichigainai*)

かねる, はず and に違いない (Tables 33, 34 and 35) are the only modal auxiliaries signalling necessity that have epistemic readings. The first one (*kaneru*), implies inability due to an unpleasant or painful reason and are used to express polite refusal and disapprovals (Narrog, 2009a, p. 98). Since it is mainly used in formal context, we cannot expect to find many examples in C-ORAL-JAPÓN. Aside from its deontic meaning of internal inability of the speaker, it can also be used to express probability in uncontrollable SOAs, rendering it epistemic in some situations. When negated, it involves a double negation, changing it into a possibility, similar to 'might' (Makino & Tsutsui, 1995, p. 98). The two remaining auxiliaries, はず (*hazu*) and に違いない (*nichigainai*), are very similar. Both involve a high degree of certainty of a SOA, resulting in an exclusive epistemic reading. The differences rely on the negation: the former can accept both propositional and auxiliary negation (with no change in modality). The latter is already a negative construction meaning 'not differing from' (違う *chigau* 'to differ' + ない *nai*, negative particle), and it cannot accept any kind of negation.

Table 33: Information for V + かねる

V + かねる	
Issue	**Result**
English Equivalents	Cannot/Could not
Modality Type	Necessity
Modality Subtype	Ambiguous
Negation Change	Yes
Negation of Auxiliary	Possible
Negation of Proposition	Not possible
Probability Percentage	0%, 50% if negated
Obligatory Tag	<m modtype="NEC" subtype="AMBG" class="AUX" value="0%">

Table 34: Information for V + はず

V + はず	
Issue	**Result**
English Equivalents	Surely / Will
Modality Type	Necessity
Modality Subtype	Epistemic
Negation Change	No
Negation of Auxiliary	Possible
Negation of Proposition	Possible
Probability Percentage	100%
Obligatory Tag	<m modtype="NEC" subtype="EPIS" class="AUX" value="100%">

Table 35: Information for V + に違いない

V + に違いない	
Issue	**Result**
English Equivalents	Surely / Will
Modality Type	Necessity
Modality Subtype	Epistemic
Negation Change	No
Negation of Auxiliary	Not possible
Negation of Proposition	Not possible
Probability Percentage	100%
Obligatory Tag	<m modtype="NEC" subtype="EPIS" class="AUX" value="100%">

V + てもいい (*temoii*)

The てもいい expression belongs to the possibility type of modality as it indicates a permission from the speaker, similar to English 'may' (Table 36). It is formed by the conditional verb ending ても (*temo*), adding the adjective auxiliary いい (*ii* 'good') meaning literally 'it is good/OK if V'. The main verb can be used with the negative, creating a sentence meaning 'it is OK if you don't V'.

Table 36: Information for V + てもいい

V + てもいい	
Issue	**Result**
English Equivalents	Can, may
Modality Type	Possibility
Modality Subtype	Deontic
Negation Change	No
Negation of Auxiliary	Not possible
Negation of Proposition	Possible
Probability Percentage	50%
Obligatory Tag	<m modtype="POSS" subtype="DEON" class="AUX" value="50%">

V + ことができる (*kotogadekiru*)

This auxiliary is similar to Spanish *Poder* + V or English *can/may* in the sense that it indicates possibility, but it only has a deontic readings (ability or capacity) (Table 37). If the direct object and the main verb have a strong relation and if the context is providing sufficient information, the main verb may be omitted as in *Poder*, as seen in the following chapter. Also, it is the only possibility marker in Japanese that changes into a necessity when negated.

The elements こと-が (*koto-ga*) are a combination of nominaliser suffix *koto* and nominative case particle *ga* which act as nominalisers of the verb of the proposition, linking it with the proper auxiliary できる (*dekiru*). However, in some cases, when the main verb is formed with the auxiliary する (*suru*), できる may replace it completely without the need of the nominalizers, or using only particle *ga*.

Table 37: Information for V + ことができる

V + ことができる	
Issue	**Result**
English Equivalents	Can, may
Modality Type	Possibility
Modality Subtype	Deontic
Negation Change	Yes
Negation of Auxiliary	Possible
Negation of Proposition	Possible
Probability Percentage	50%, 0% if negated
Obligatory Tag	<m modtype="POSS" subtype="DEON" class="AUX" value="50%">

V + かもしれない (*kamoshirenai*)

かも知れない is the standard auxiliary for signalling epistemic possibility in Japanese, and the most frequent one. It is already negated but it does accept negative propositions (Table 38).

Table 38: Information for V + かもしれない

V + かもしれない	
Issue	**Result**
English Equivalents	Can, may
Modality Type	Possibility
Modality Subtype	Epistemic
Negation Change	No
Negation of Auxiliary	Not possible
Negation of Proposition	Possible
Probability Percentage	70%
Obligatory Tag	<m modtype="POSS" subtype="EPIS" class="AUX" value="70%">

V + とは限らない (*tohakagiranai*) and **V + ほどのこともない** (*hodonokotomonai*)

These two very similar expressions state that the SOA is not necessarily true. The auxiliaries are already in the negative form (**ない**), so they do not accept additional negative elements. They will be tagged as possibility since the event is open to happen (Tables 39 and 40).

Table 39: Information for V + とは限らない

V + とは限らない	
Issue	**Result**
English Equivalents	No need
Modality Type	Possibility
Modality Subtype	Epistemic
Negation Change	No
Negation of Auxiliary	Not possible
Negation of Proposition	Not possible
Probability Percentage	50%
Obligatory Tag	<m modtype="POSS" subtype="EPIS" class="AUX" value="50%">

Table 40: Information for V + ほどのこともない

V + ほどのこともない	
Issue	**Result**
English Equivalents	No need
Modality Type	Possibility
Modality Subtype	Epistemic
Negation Change	No
Negation of Auxiliary	Not possible
Negation of Proposition	Not possible
Probability Percentage	50%
Obligatory Tag	<m modtype="POSS" subtype="EPIS" class="AUX" value="50%">

V + だろう (*darō*)

This auxiliary is somewhat problematic regarding its great variety of meanings and special treatment it has received by many experts. It has a standard, possible epistemic reading, indicating a possibility of a SOA perceived by the speaker, similar to the previous かもしれない (Table 41). However, its meaning changes in interrogative contexts, when the speaker asks for information or or confirmation of an information he has previously expressed. The only だろう considered for this study as a modal marker will be the first interpretation, when it is used in an enunciative, non-interrogative sentence.

Table 41: Information for V + だろう

V + だろう	
Issue	**Result**
English Equivalents	Can, may
Modality Type	Possibility
Modality Subtype	Ambiguous
Negation Change	No
Negation of Auxiliary	Not possible
Negation of Proposition	Possible
Probability Percentage	70%
Obligatory Tag	<m modtype="POSS" subtype="EPIS" class="AUX" value="70%">

To finish with the auxiliaries, the next table, Table 42, shows a comparison between Japanese forms and their equivalent in Spanish and English, as well as the subtype of modality (Epistemic, Deontic or both) that is implied.

Table 42: Japanese and Spanish auxiliaries with their English equivalents and modality subtype

Japanese	**Spanish**	**English**	**Subtype**
V + なければならない (*nakerebanaranai*)	*Tener, Haber que / Deber* + V	Have to / Must	Deontic
V + ざるを得ない (*zaruwoenai*)	*Tener, Haber que / Deber* + V	Have to / Must	Deontic
V + しかない (*shikanai*)	*Tener, Haber que / Deber* + V	Have to / Must	Deontic
V + 訳にはいかない (*wakenihaikanai*)	*No* + *Poder* / *Deber* + V	Cannot / Must not	Deontic
V + に忍びない (*nishinobinai*)	*No* + *Poder* + V	Cannot / Could not	Deontic
V + べき (*beki*)	*Tener, Haber que / Deber* + V	Should / Ought to	Deontic
V + た方がいい (*tahōgaii*)	*Tener, Haber que / Deber* + V	Should / Ought to	Deontic
V + たらいい (*taraii*)	*Tener, Haber que / Deber* + V	Should / Ought to	Deontic
V + ればいい (*rebaii*)	*Tener, Haber que / Deber* + V	Should / Ought to	Deontic
V + たい (*tai*)	∅[15]	∅	∅

[15]No grammaticalised equivalent

V + てもらいたい (*temoraitai*)	∅	∅	∅
V + ほしい (*hoshii*)	∅	∅	∅
V + ください (*kudasai*)	∅	∅	∅
V + つもり (*tsumori*)	*Ir a* + V	Will / Going to	Deontic
V + かねる (*kaneru*)	*No* + *Poder* + V	Cannot / Could not	Ambiguous
V + はず (*hazu*)	*Ir a* + V	Will / Going to	Epistemic
V + に違いない (*nichigainai*)	*Ir a* + V	Will / Going to	Epistemic
V + てもいい (*temoii*)	*Poder* + V	Can / May	Deontic
V + ことができる (*kotogadekiru*)	*Poder* + V	Can / May	Deontic
V + かもしれない (*kamoshirenai*)	*Poder* + V	Can / May	Epistemic
V + とは限らない (*tohakagiranai*)	*No* + *Tener, Haber que* + V	Not have to	Epistemic
V + ほどのこともない (*hodonokotomonai*)	*No* + *Tener, Haber que* + V	Not have to	Epistemic
V + だろう (*darō*)	Poder + V	Can / May	Epistemic

The larger array of different modal auxiliaries for Japanese rapidly contrasts with the limited one among the Spanish and English counterparts. Spanish auxiliaries are forced to contain more different meanings that inevitably lead to more ambiguity. The next chapter, focus on their presence in the corpora, will provide

an insight on this matter. Also, the Japanese auxiliaries related to the speaker's desires and petition do not have their grammaticalised counterpart in English and Japanese. They could be translated as *querer* or *want*, but since these have been ruled out as 'semiauxiliaries', they are not included in this study.

3.5.2 Adverbs and adjectives

Adverbs and adjectives involve less complications than the auxiliaries since they are only formed by a single word expression. As explained in Chapter 2, they can be negated either with an independent morpheme, or through a negative prefix. Also, there seems to be a series of semantic or scope restrictions in terms of negation as not all of them can be modified by a negative element. Those than can be negated will necessarily change the modality type to the opposite one. For example, Spanish *necesariamente* ('necessarily') can be negated as in *no necesariamente* ('not necessarily'), becoming a possibility. On the other hand, an adverb like *seguramente* ('surely') cannot be negated as **no seguramente* or **inseguramente* ('not surely'), and requires a different adverb to encode that meaning.

In relation to the type of modality, both necessity and possibility can be marked in an adverb or adjective in both languages, but only epistemic modality. An adverb or an adjective will only address the probability of the state of affairs of becoming true, and cannot interact with the receiver of the message to achieve it.

3.5.2.1 Spanish adverbs and adjectives

The following tables display the possible modal adverbs (Tables 43 and 44) and adjectives (Tables 45 and 46) that can encode modality in Spanish, and their corresponding obligatory tag.

Among all adverbs, some of them are formed by attaching the negative prefix *in-* (or *im-* if followed by letters *b* or *p*) marking the opposite. Necessity adverbs *imposiblemente, improbablemente, indubitablemente, indubitadamente, indudablemente, ineludiblemente, inevitablemente* and *innegablemente* are formed from

Table 43: Spanish necessity adverbs

Adverb	Engish translation	Obligatory tag
Seguramente	Surely, certainly	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Categóricamente	Decisively	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Certificadamente	Surely, certainly	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Ciertamente	Surely, certainly	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Indefectiblemente	Unfailingly	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Decididamente	Definitely	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Definitivamente	Definitely	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Forzadamente	Necessarily	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Forzosamente	Necessarily	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Imposiblemente	Impossibly	<m modtype="NEC" subtype="EPIS" class="Adverb" value="0%">
Improbablemente	Improbably	<m modtype="NEC" subtype="EPIS" class="Adverb" value="0%">
Indubitablemente	Undoubtedly	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Indudablemente	Undoubtedly	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Ineludiblemente	Inevitably	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Inevitablemente	Inevitably	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Innegablemente	Undeniably, irrefutably	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Necesariamente	Necessarily	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Obviamente	Obviously	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Obligatoriamente	Obviously	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Palpablemente	Obviously	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Sin discusión	Without discussion	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Sin duda	Without doubt	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
Sin falta	By all means	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">

Table 44: Information for Spanish possibility adverbs

Adverb	**Engish translation**	**Obligatory tag**
Quizá(s)	Perhaps	<m modtype="POSS" subtype="EPIS" class="Adverb" value="70%">
A lo mejor	Perhaps	<m modtype="POSS" subtype="EPIS" class="Adverb" value="70%">
Probablemente	Probably	<m modtype="POSS" subtype="EPIS" class="Adverb" value="70%">
Posiblemente	Possibly	<m modtype="POSS" subtype="EPIS" class="Adverb" value="70%">
Innecesariamente	Unnecessarily	<m modtype="POSS" subtype="EPIS" class="Adverb" value="50%">
Inciertamente	Uncertainly	<m modtype="POSS" subtype="EPIS" class="Adverb" value="50%">
Accidentalmente	Accidentally	<m modtype="POSS" subtype="EPIS" class="Adverb" value="30%">
Difícilmente	Uncertainly	<m modtype="POSS" subtype="EPIS" class="Adverb" value="30%">
Dudosamente	With doubt	<m modtype="POSS" subtype="EPIS" class="Adverb" value="30%">
Extrañamente	With doubt	<m modtype="POSS" subtype="EPIS" class="Adverb" value="30%">
Hipotéticamente	Hypothetically	<m modtype="POSS" subtype="EPIS" class="Adverb" value="30%">
Incidentemente	Accidentally	<m modtype="POSS" subtype="EPIS" class="Adverb" value="30%">
Incidentalmente	Accidentally	<m modtype="POSS" subtype="EPIS" class="Adverb" value="30%">

their possibility counterparts. However, some of then, (*ineludiblemente, indubitablemente, indubitadamente*) do not exist without the negative prefix and do not appear in Table 44 (**dubitablemente*, **dubitadamente*, **dudablemente*, **eludiblemente*, **innegablemente*). Their adjective form is grammatical (*dubitable, dubitado, dudable, eludible*), which proves the negative conversion was made in the adjective form, before becoming an adverb with the suffix -mente. The opposite also takes place: possibility adverbs are formed by adding the negative prefix to a necessity word (*innecesariamente* and *inciertamente*). Tables 45 and 46 represent the adjectives from which these adverbs are formed.

Table 45: Spanish necessity adjectives

Adjective	**Engish translation**	**Obligatory tag**
Seguro	Sure, certain	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">
Categórico	Decisive	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">
Definitivo	Definite	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">
Imposible	Impossible	<m modtype="NEC" subtype="EPIS" class="Adjective" value="0%">
Improbable	Improbable	<m modtype="NEC" subtype="EPIS" class="Adjective" value="0%">
Cierto	Sure, certain	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">
Indefectible	Unfailing	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">
Indubitable	Undoubted	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">
Indubitado	Undoubted	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">
Indudable	Undoubted	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">
Ineludible	Inevitable	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">
Inevitable	Inevitable	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">
Innegable	Undeniable, irrefutable	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">
Necesario	Necessary	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">
Obvio	Obvious	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">
Obligatorio	Obvious	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">

The tables show that not every adjective maintains the modal meaning when transformed into an adverb by the suffix -*mente* and vice-versa. For example, adverbs *certificadamente, decididamente, forzosamente, forzadamente, palpablemente, difícilmente, extrañamente, incidentemente* and *incidentalmente* lose their modal meaning without suffix -*mente*. Also, I have considered a group of modal adverbs that are not formed with adverb -*mente* from an adjective, which are in most cases multiword adverbs, such as *sin discusión, sin duda, sin falta, quizá(s)* and *a lo mejor*.

Table 46: Spanish possibility adjectives

Adjective	Engish translation	Obligatory tag
Probable	Probable	<m modtype="POSS" subtype="EPIS" class="Adjective" value="70%">
Posible	Possible	<m modtype="POSS" subtype="EPIS" class="Adjective" value="70%">
Innecesario	Unnecessary	<m modtype="POSS" subtype="EPIS" class="Adjective" value="50%">
Incierto	Uncertain	<m modtype="POSS" subtype="EPIS" class="Adjective" value="50%">
Accidental	Accidental	<m modtype="POSS" subtype="EPIS" class="Adjective" value="30%">
Dudoso	Uncertain	<m modtype="POSS" subtype="EPIS" class="Adjective" value="30%">
Hipotético	Hypothetical	<m modtype="POSS" subtype="EPIS" class="Adjective" value="30%">

3.5.2.2 Japanese adverbs and adjectives

Tables 47 and 48, 49 and 50 represent the possible modal adverbs and adjectives in Japanese, and their corresponding obligatory tag.

Table 47: Japanese necessity adverbs

Adverb	Engish translation	Obligatory tag
是非	(*zehi*) Surely, certainly	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
必ず	(*kanarazu*) Necessarily, certainly	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
絶対 (に)	(*zettai*) Absolutely, unconditionally	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
嘸 (かし/や)	(*sazo(kashi/ya)*) Certainly, surely	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
屹度	(*kitto*) Undoubtedly	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
確かに	(*tashikani*) Certainly, surely	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">
相違なく	(*sōinaku*) Certainly, surely	<m modtype="NEC" subtype="EPIS" class="Adverb" value="100%">

Table 48: Japanese possibility adverbs

Adverb	Engish translation	Obligatory tag
多分	(*tabun*) Surely, certainly	<m modtype="POSS" subtype="EPIS" class="Adverb" value="70%">
恐らく	(*osoraku*) Perhaps, probably	<m modtype="POSS" subtype="EPIS" class="Adverb" value="70%">
或いは	(*aruiha*) Possibly	<m modtype="POSS" subtype="EPIS" class="Adverb" value="50%">
若しか (すれば/して)	(*moshika(sureba/shite)*) Possibly	<m modtype="POSS" subtype="EPIS" class="Adverb" value="50%">
ひょっとしたら	(*kitto*) Possibly, perhaps	<m modtype="POSS" subtype="EPIS" class="Adverb" value="50%">

Three adverbs are formed directly from adjectives. 絶対に and 確かに have been formed adding particle に (*ni*) to their respective nominal adjectives, and 相違なく is the く (*ku*) form of the 'pure' adjective 相違ない (*sōinai*, 'certain, sure'). The following tables show that adjective 可能 (*kanō*, 'possible') is the only possibility adjective, and its contrary ('not possible') is formed by adding the negative prefix kanji 不 (不可能 (*fukanō*).

This section has described the adverbs and adjectives that encode modality in Spanish and Japanese. There are less modal adverbs and adjectives available in Japanese than in Spanish, with only one fully grammaticalised adjective to signal possibility in a predicative position. Whereas Japanese has a much greater number of deontic and auxiliary markers than the latter, Spanish stands out with a higher variety of epistemic elements. The corpus study will reveal if the number of different markers available is correlated with their frequency. The next and final subsection will tackle the last modal marker, the mood marking of modality.

Table 49: Japanese necessity adjectives

Nominal adjective	**Engish translation**	**Obligatory tag**
確か (な)	(*tashika*) Sure, certain	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">
確実 (な)	(*kakujitsu*) Sure, certain	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">
駄目 (な)	(*dame*) Can/Must not	<m modtype="NEC" subtype="EPIS" class="Adjective" value="0%">
インポッシブル (な)	(*imposhiburu*) Impossible	<m modtype="NEC" subtype="EPIS" class="Adjective" value="0%">
必要 (な)	(*hitsuyō*) Necessary	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">
必須 (な)	(*hissu*) Necessary, indispensable	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">
必要不可欠 (な)	(*hitsuyōfukaketsu*) Necessary, imperative	<m modtype="NEC" subtype="EPIS" class="Adjective" value="100%">
無理 (な)	(*muri*) Impossible	<m modtype="NEC" subtype="EPIS" class="Adjective" value="0%">
不能 (な)	(*funō*) Impossible	<m modtype="NEC" subtype="EPIS" class="Adjective" value="0%">
不可能 (な)	(*fukanō*) Impossible	<m modtype="NEC" subtype="EPIS" class="Adjective" value="0%">
確実 (な)	(*dame*) Can/Must not	<m modtype="NEC" subtype="EPIS" class="Adjective" value="0%">

Table 50: Japanese possibility adjective

Nominal adjective	**Engish translation**	**Obligatory tag**
可能 (な)	(*kanō*) Possible	<m modtype="POSS" subtype="EPIS" class="Adjective" value="50%">

3.5.3 Verbal mood

As described before, verbal moods have progressively lost their specific meaning. Whereas it can be argued that indicative and subjunctive moods indicate realis or irrealis meanings, they are still very vague and require additional auxiliaries to express specific modal senses. The only mood that appears to have retained a more specific use in both Spanish and Japanese would be the imperative mood, shown in Table 51, and the potential mood in Japanese. The tagging will be the same for both languages, with a small exception: the class attribute for the Spanish negative imperative will feature as a 'subjunctive', since it is formed by a negative element and the verb in subjunctive mood.

Table 51: Imperative forms in Spanish and Japanese

Language	**Example**	**Engish**	**Obligatory tag**
Spanish affirmative	Ven rápido!	Come quickly!	<m modtype="NEC" neg="no" subtype="DEON" class="mood_IMP" value="100%">
Japanese affirmative	早く来い！		<m modtype="NEC" neg="no" subtype="DEON" class="mood_IMP" value="100%">
Spanish negative	No te comas eso!	Don't eat that!	<m modtype="NEC" neg="yes" subtype="DEON" class="mood_SUBJ" value="0%">
Japanese negative	それを食べるな！		<m modtype="NEC" neg="yes" subtype="DEON" class="mood_IMP" value="0%">

This ends Chapter 3 of the study. These pages have explained the methodology of the study, the tools, and the tagset for each possible marker. The study is structured in four major steps: (1) the theoretical foundation, explained in Chapter 2; (2) the preparation of the corpora and XML tagset; (3) the annotation of modality in the corpora, following the guidelines of (1) and (2) ; and finally (4), the creation of an automatic tagger, which has been designed using the knowledge from the corpora (3) and the tagset (1).

The tools used for the study are two corpora, C-ORAL-ROM and C-ORAL-

JAPÓN, one for each language, a mark-up language (XML), a programming language (Python) and two POS taggers, one Spanish, Grampal, and one Japanese, Juman.

Finally, we have listed each possible modal markers that may appear in the corpora. Each marker has been analysed, outlining their major idiosyncratic characteristics and compared in both languages, and their corresponding XML tag was included. Two main conclusions can be extracted from this section: first, there is a very different variety of modal markers in both languages, as represented in Table 52. The corpus study will also show if a higher number of available markers is correlated with their usage frequency. Second, auxiliaries have proven to be the most ambiguous. Since there are less Spanish markers available, only seven, we should expect a high amount of ambiguity in the corpus. The next Chapter will cover the results extracted from the annotation of the corpora, providing statistical information about the type of modality and markers found in each corpus.

Table 52: Modal markers available for each language

Class	**Modality**	**Spanish**	**Japanese**
Auxiliaries	Necessity [Epistemic]	1	2
	Necessity [Deontic]	3	14
	Necessity [Ambiguous]	2	1
	Possibility [Epistemic]	0	9
	Possibility [Deontic]	0	9
	Possibility [Ambiguous]	1	0
Adverbs	Necessity [Epistemic]	23	7
	Possibility [Epistemic]	13	5
Adjectives	Necessity [Epistemic]	16	11
	Possibility [Epistemic]	7	1
Mood	Necessity [Deontic]	1(2)	2

Chapter 4

Corpus Study. Results and discussion.

4.1 Preliminary hypotheses

This chapter will describe through a collection of tables and graphs the results of the statistical analyses about the modality found in C-ORAL-ROM and C-ORAL-JAPÓN corpora. The main objective is to observe frequency and distribution of patterns in modal markers among the speakers, if there is a regularity in their usage or if it is otherwise random, and how it differs according to language. The study will take into the account linguistic factors of language, type of discourse and register; non-linguistic, like gender and age of the participants; and elements of the sentence that could modify them such as negation, ellipsis and syntactic separation.

As preliminary hypotheses, I believe the results may show several facts:

1. There are not any known previous records of cross-linguistic quantitative studies between Spanish and Japanese modality to compare to, and the general results regarding its usage are unknown. The main question to answer in this section is how both languages differ quantitatively in terms of modal markers. If modality is a universal feature encoded in every language such as tense or aspect, there should not be a wide difference in terms of overall numbers. In other words, when observing modality in general in the corpora, we consider as the first null hypothesis that the usage in both languages is not related and therefore the differences would be considerable. The alternative hypothesis is that both means are going to be related.
2. We could expect however necessity modality to be higher in Spanish than in Japanese. Discourse studies have shown the differences between Asian and Western languages involving politeness and face threatening acts (FTAs). It is believed that Japanese avoids direct statements from the speaker in order preserve the face or respect towards the receiver of the message (Matsumoto, 1989; Ide, 1992). For this reason, necessity markers, since they involve an absolute certainty or directness, may not be used by Japanese speakers. Spanish, on the other hand, may use these markers more freely. To summarise, the second hypothesis would be that the means between Spanish and Japanese necessity are significantly different.

3. In terms of subtype of modality, following the same argumentation, deontic markers, those that impose a necessity or possibility upon the recipient of the message to achieve a state of affairs, could be less frequent in Japanese. This leads us to the third hypothesis, that the difference between Spanish and Japanese deontics will be significant. Also, regarding ambiguity, we have seen in Subsection 3.5.1 that Japanese uses a higher number of markers depending on the situation, which makes it less probable to find ambiguous markers than in Spanish.
4. In terms of selection of modal markers, our references (Gómez Manzano, 1991; Narrog, 2009a) only consider auxiliaries (periphrases, suffixes) in the corpora. Among the Spanish ones, *poder* + V ('can', 'may'), *ir a* + V ('will', 'going to') and *tener que* + V ('have to', 'should') (p. 213) appear to be the most frequent ones. In Japanese (p.167), たい (*tai*, 'want to'), なければならない (*nakerebanaranai*, 'have to') and ことができる (*kotogadekiru* 'can') seem to achieve the highest numbers. The fourth hypothesis, then, would be that there is a significant difference in the usage of auxiliaries.
5. Modality will also be influenced by the type of discourse (monologue vs dialogue/conversation). Since it involves the probability of the state of affairs of the speaker in relation to the receiver of the message, I believe modality in general will be significantly higher in interactive situations (dialogue/conversations) than monologues.
6. Modality is related to the level of register (informal vs formal). Preliminary studies (Herrero & Moreno, 2014) have shown that modality is higher in formal monologues than in informal ones. However, my hypothesis is that in interactive situations, informal modals will be higher.
7. Differences in use in gender and age factors will be higher in Japanese than in Spanish, and will affect the choice of modality. There is a higher variation in Japanese, especially among verb inflection, auxiliaries and particles in the spoken discourse, according to the gender, age, social position and region accent of the participants. C-ORAL-JAPÓN does not have enough variety of speakers in terms of social position and accent, but we believe the results regarding gender and age will be sufficient to draw some conclusions. Gómez's

study (1991, p. 214) shows that younger speaker use more modal auxiliaries than older people. My study will observe if this is the same case in C-ORAL-ROM.

8. Modification of markers will also be taken into account in both languages, mainly negation, ellipsis and separation of auxiliaries. First, negation is a universal principal of human language (Horn, 1989), and fairly frequent issue in spoken language (Biber et al., 1999; Herrero, 2013b), so we can safely assume to find a higher amount of negated modals in both languages. Secondly, ellipsis is also a universal feature (Gilligan, 1987), but its frequency is not the same in each language. For example, whereas in English is barely present, in Japanese is almost constant, with 70% of nominal arguments omitted in spoken discourse, leaving in most cases only the verb in the sentence (Nariyama, 2003). Our hypothesis in this matter is that we may find higher instances of modal auxiliaries ellipsis in the Japanese corpus than the Spanish one. However, since the auxiliary is attached to a main verb, which is not so frequently omitted, we believe the proportion would not be so high. Thirdly and finally, separation of the auxiliary and the main verb can appear in the corpora, but the distance may be higher in Spanish since the auxiliary is an independent word.

In order to validate these hypotheses we will observe the frequencies of the data per each speaker of the corpus. First by the total number of modal markers, then the type of modality, subtype and their grammatical class. The frequencies have been normalised to 1000 words. The reason to do this is to obtain a relative view of the usage of modal markers according to the amount of words uttered by a speaker. A higher amount of markers may or may not be directly associated with a high number of words. A normalised frequency takes into account a number in relation to the totality of the data (McEnery & Hardie, 2012). This allows us to ensure the according proportional numbers and make proper comparisons intra and inter-linguistically.

The normalisation takes into account the number of words uttered by each speaker. For this, the complete transcribed text of each speaker has been word-

tokenised and counted. For the Japanese corpus, Juman was used. This may be problematic, as the Japanese tokeniser may *oversegment* some words, specially auxiliaries and verb inflective morphemes and make an erroneous counting of the words. However, the use of a tokeniser or tagger is essential for word-counting a Japanese text as it would be nearly impossible to do by hand, which forces us to use the frequency count made by Juman.

Regarding the quantitative analysis, it is mainly formed by descriptive statistics representing the frequency distributions in the corpora. Inferential statistics and tests will be used to confirm the previous hypotheses.

The chapter is divided as followed: Section 4.2 will describe in general the overall numbers of markers and modality type among the speakers of both corpora; Section 4.3 observes the usage of modality according to the type of register and discourse; Section 4.4 will study the differences of usage according to age and gender; Section 4.5 shows the frequencies of each marker; and Section 4.6 will discuss factors that modify the markers, such as negation, ellipsis or separation.

4.2 Modality among corpora

4.2.1 Modality: General numbers

The following Table (53), shows the number of files, speakers, words and mean of words per speaker, and the total amount of modal markers (absolute and normalised per 1000 words frequencies). The complete frequencies can be found in Appendix A.

Table 53: Files, speakers, words and modal markers in both corpora

Corpus	**Files**	**Women**	**Men**	**Words(μ/speakr)**	**Mod.Mrkrs(norm.)**
CORAL-ROM	169[1]	154	225	301,329(795.06)	3951(12.84)
CORAL-JP	39	37	21	127,676(2201.31)	1076(8.43)

The number of files, speakers and words is considerably lower in C-ORAL-JAPÓN, with four and more than ten times less of men and women than C-ORAL-ROM respectively, and nearly a third of the amount of words. However, the higher mean of words per speaker indicates longer speaking periods. Regarding the amount of modal markers, the normalised frequencies show higher numbers, exactly 1.52 times more, in C-ORAL-ROM (12.84) than in C-ORAL-JAPÓN (8.43).

In the following pages we will take a closer look on this difference among the speakers, with all the frequencies normalised by 1000 words. To begin with, Figure 11 shows how the mean of modal markers per speaker differs in size, and Figure 12 followed by Table 54, the general calculations and dispersion between Spanish and Japanese modality as a whole.

The frequencies show a higher amount of markers in Spanish. However, the minimum and maximum dispersion is quite small and similar between both languages. The majority of markers are situated around 4 in the 25%, and 11 and 17 in Spanish and Japanese respective 75% percentiles. In other words, both languages

[1]The C-ORAL-ROM telephone conversations files have not been included in the study (see `http://www.lllf.uam.es/ING/Datos_Coralrom.html`)

Figure 11: Modal markers per speaker (mean) in Spanish and Japanese

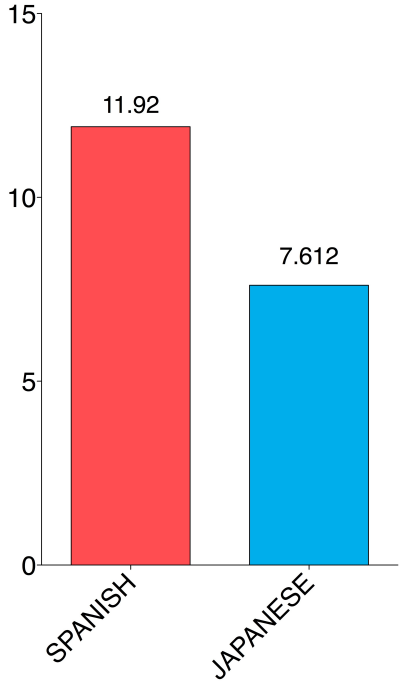
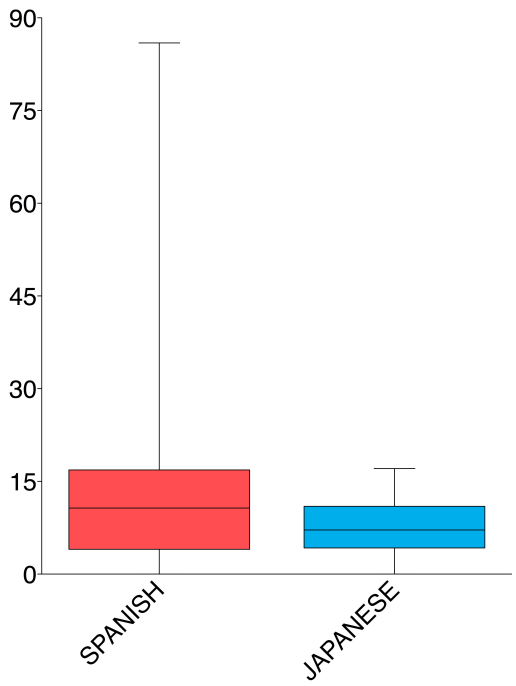


Figure 12: Dispersion of modal markers in Spanish and Japanese corpora



are similar in the majority usage of markers, especially among the speakers that use less modal markers. The Japanese, however, are much more concentrated around the median, with a SD of 4.69 compared to the Spanish SD of 10.67.

The differences are higher observing the dispersion of the data. Regarding maximums and minimums, there is an abnormal maximum point of 85.94 markers in the Spanish corpus, represented by the abnormally large *whisker* in Figure 12. Looking at the data, this belongs to speaker 'MAM_SPAwoman_C_ Segovia', which has used 10 instances of necessity-deontic modality (verbs in imperative mood to be precise), and has only spoken 128 words. The conversation that takes place at a household, and this speaker is interrupting a longer dialogue between two participants, constantly issuing them orders, without taking part in the interaction, rendering it an isolated case. It will, nevertheless, affect the dispersion of the data.

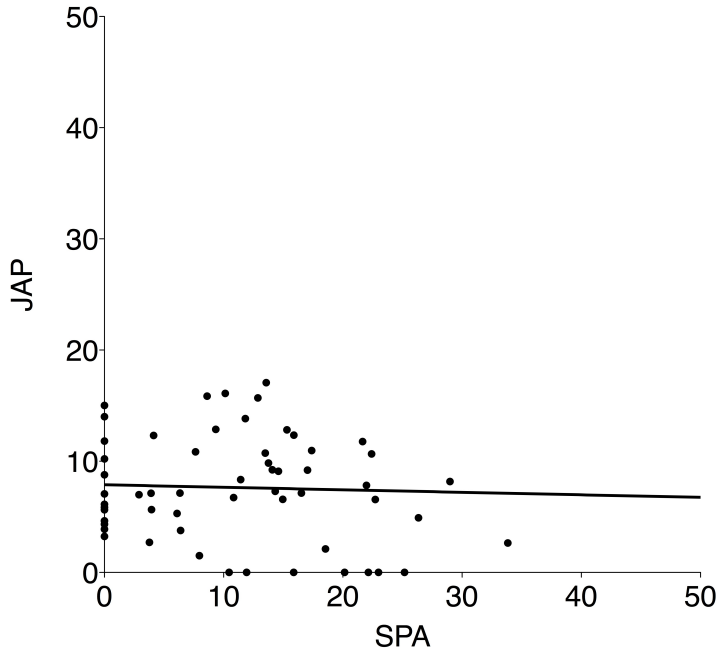
Table 54: Column statistics for Spanish and Japanese modality

Stat	**Spanish**	**Japanese**
N[2]	379	58
Minimum	0.0	0.0
25% Percn.	4.016	4.228
Median	10.70	7.142
75% Percn.	16.85	10.98
Maximum	85.94	17.06
Mean	11.92	7.612
SD	10.67	4.695
KS test		
P value	< 0.0001	> 0.10
Passed?	No	Yes
P value summary	***	ns
D'Agostino and Pearson test		
P value	< 0.0001	0.2623
Passed?	No	Yes
P value summary	***	ns
Shapiro-Wilk test		
P value	< 0.0001	0.1354
Passed?	No	Yes
P value summary	***	ns

[2]Number of values (speakers)

If we trace a linear regression of the data (Figure 13) situating the Japanese numbers on the Y axis and the Spanish on the X axis, we can see a higher concentration of points around the 10-15 values, and an almost horizontal line with a slight negative slope. This indicates that, as the number of the markers per speaker increases in Spanish, the average quantity remains the same in Japanese.

Figure 13: Linear regression of Spanish vs Japanese modal markers



Furthermore, a t test shows the difference between the means is significantly different, rejecting the idea of a similar relation in terms of modality usage and the alternative hypothesis and accepting the null hypothesis ($P < 0.05$) (Table 55).

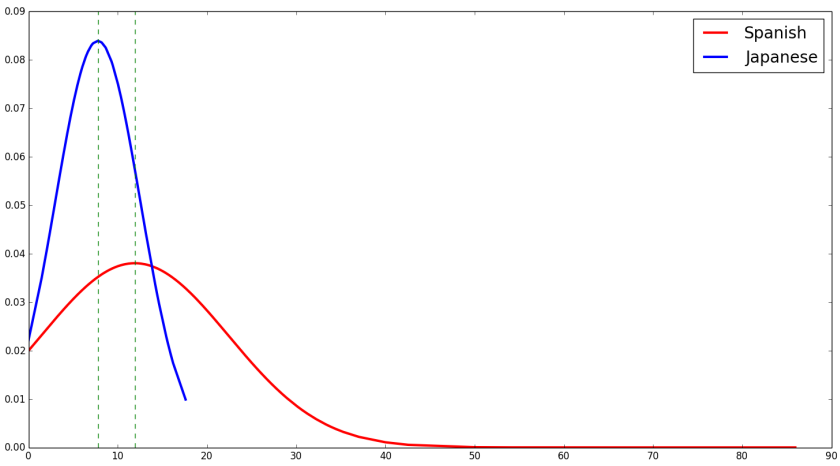
Table 55: Mann Whitney test (two-tailed) results for Spanish and Japanese modality

Feature	**Result**
P value	0.0033
P summary	**
Signf. different	Yes

Lastly, looking at the probability distribution, three different normality tests were performed on the data to check if they have been sampled from a standard

normal population. In order words, verify how closely the majority of the data is concentrated symmetrically around the mean. As seen in Table 54, the three tests, KS, D'Agostino and Pearson and Shapiro-Wilk tests have resulted negative for Spanish (with a p value of < 0.0001), but positive for Japanese (p values of $>$ 0.10, 0.26 and 0.13, respectively). Figure 14 represents these as gaussian curves. The higher and sharper curve in Japanese shows a higher concentration of data around the mean and a higher probability than Spanish. Spanish data has a higher variation of data which is less probable than Japanese.

Figure 14: Probability distribution of modal markers in Spanish and Japanese corpora

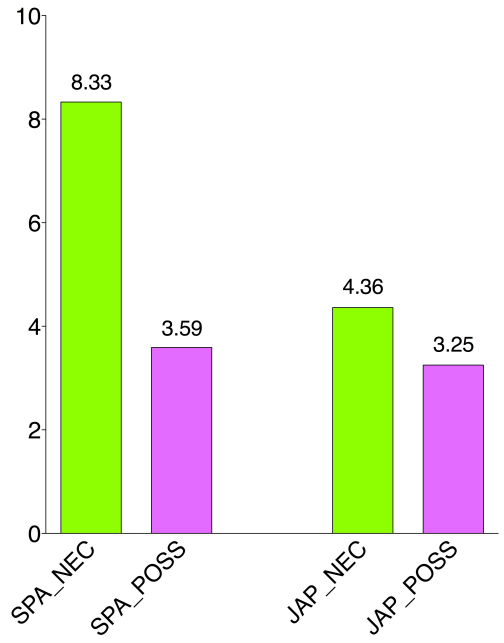


In conclusion, Spanish speakers use a 1.5 higher amount average of modal markers than Japanese. The difference between the means is sufficiently higher to accept the null hypothesis, the usage of modality in both languages is significantly different and not related. The markers in the Spanish corpus show higher irregularities due to abnormal maximum values and fail normality tests. In the Japanese corpus, however, modality appears to be much more concentrated and regular, passes normality tests, and therefore is highly predictable. Nevertheless, there are similarities among the values inside both standard evaluations: the majority of speakers use similar amounts of modal markers. A closer look to the type of modality used by each speaker will help to clarify this.

4.2.2 Necessity vs possibility

The following Figures show the results of the same analyses but comparing the uses of necessity and possibility markers. First, looking at the total means (Figure 15), in Spanish speakers necessity is 2.3 higher than possibility modality, whereas in Japanese the difference is only 1.3. The mean per speaker of necessity markers is around two times higher in Spanish than Japanese, confirming our second assumption of necessity been used more freely in Spanish.

Figure 15: Mean per speaker of necessity and possibility markers used in the corpora



The t tests reinforce this assumption (Table 56): there is a significant difference between Spanish and Japanese necessity modality markers, and also between Spanish necessity and possibility. However, in terms of possibility, though, the difference between the usage of Spanish and Japanese is not significant. The difference of modality seen in the previous sectionm then, appears to be caused by differences in necessity, since the usage of possibility is nearly equal.

Table 56: Mann Whitney test (two-tailed) results for necessity and possibility modality

Comparison	**Feature**	**Result**
SPA - JAP necessity	P value	0.0024
	P summary	**
	Signf. different	Yes
SPA - JAP possibility	P value	0.2619
	P summary	ns
	Signf. different	No
Necessity - possibility (SPA)	P value	< 0.0001
	P summary	***
	Signf. different	Yes
Necessity - possibility (JAP)	P value	0.0264
	P summary	*
	Signf. different	Yes

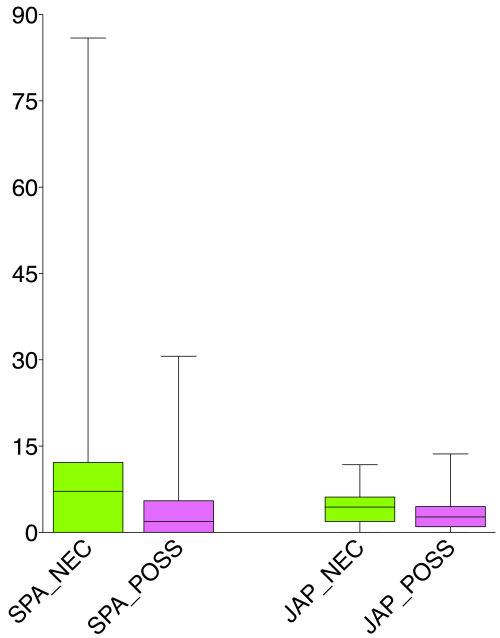
In general, the dispersion (Table 57 and Figure 16) shows a higher concentrated data around the median in Japanese necessity and possibility (SD of 3), and Spanish possibility (around 5). Necessity markers are less concentrated in the Spanish corpus, with a standard deviation of 9.3.

The data in the percentiles differs especially in the 25%, with a value of 0 in Spanish modality and more than 1 marker in Japanese. The 75% is quite similar in both languages, between 5 and 6, with the exception of Spanish necessity, that reaches 12 markers. Regarding the maximum value, once again we find lower, higher concentrated values in Japanese. Spanish possibility has a maximum point of 30, but the abnormal maximum can be found in the necessity modality (nearly 86).

Table 57: Column statistics of necessity vs possibility in Spanish and Japanese corpora

	Spanish		Japanese	
Stat	**NEC**	**POSS**	**NEC**	**POSS**
N	379	379	58	58
Min.	0.0	0.0	0.0	0.0
25% Percn.	0.0	0.0	1.907	0.9826
Median	7.156	1.912	4.410	2.686
75% Percn.	12.17	5.495	6.149	4.503
Max.	85.94	30.61	11.76	13.65
Mean	8.331	3.591	4.361	3.251
SD	9.365	4.838	3.032	2.982

Figure 16: Dispersion of necessity and possibility markers used in the corpora

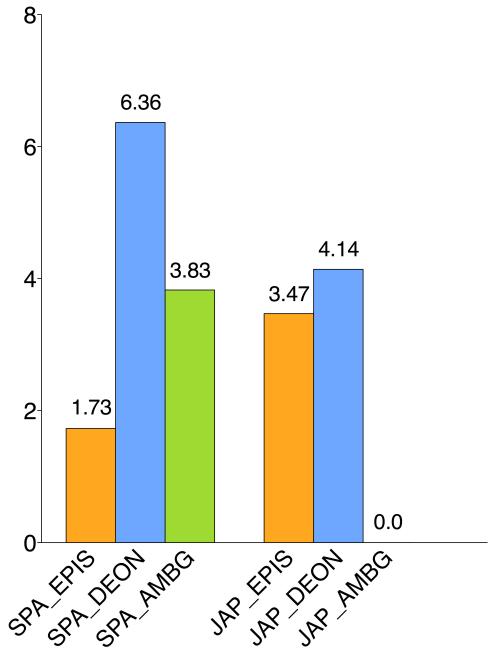


In conclusion, the main difference between Spanish and Japanese modality is located in the Spanish necessity, which has higher maximum points, and less concentration than possibility, and Japanese modality. The difference between the mean usage of necessity markers between both languages is significant, confirming our second hypothesis. Possibility markers, on the other hand, are not significantly different. Necessity is certainly the feature that most differs in each language.

4.2.3 Epistemic vs deontic modal markers

Moving now to the second level of classification of modality, the total means in Figure 17 show the different results in both languages. In both languages the main subtype of modality used is deontic, with means of 6.36 and 4.14. However, there are two main differences: first, the variance from epistemic is higher in Spanish, nearly 4 times, whereas in Japanese it is only 1.2. Secondly, there is a high amount of ambiguity in Spanish, nearly double the amount of epistemic markers, whereas there is not any in Japanese. Recalling what we described in Section 2.2.2, these markers can signal either epistemic or deontic modality, or both at the same time, since their ambiguity seems to be syntactic. The high amount of ambiguity confirms our assumption first drafted in Section 3.5.1: the small variety among Spanish modals leads to an overlap of meaning and, hence, more ambiguity.

Figure 17: Mean per speaker of epistemic, deontic and ambiguous markers in Spanish and Japanese corpora



The third hypothesis, which assumes that epistemic markers in Japanese would be significantly higher than deontic ones, appears to be false. However, the difference is not significant, like in the case of Spanish, as shown by the t test (Table 58).

Table 58: Unpaired t test (two-tailed) results for Japanese epistemic and deontic modality

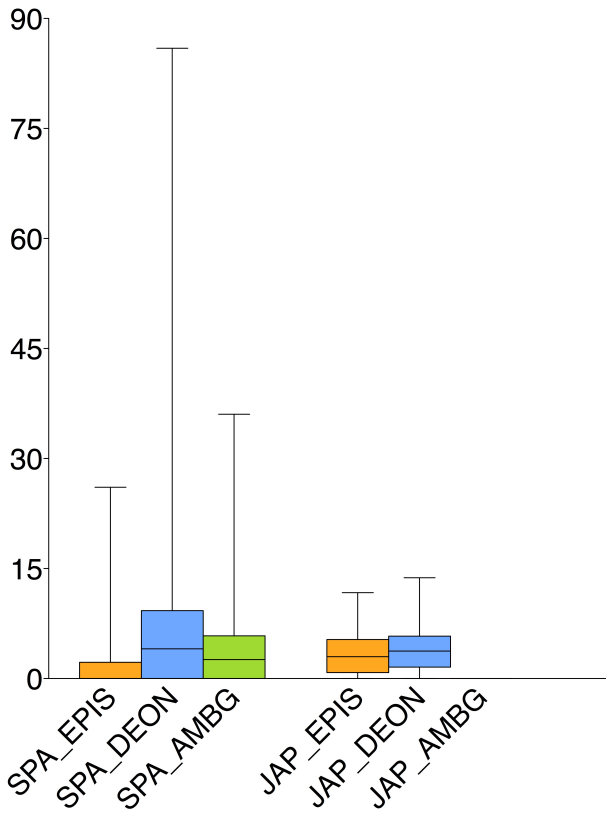
Comparison	**Feature**	**Result**
Epistemic - Deontic (JAP)	P value	0.2555
	P summary	ns
	Signf. different	No

Observing the dispersion data from Table 59 and Figure 18, the data in the Spanish corpus is more varied, with higher ranges from the median than Japanese, and also higher maximum values. The highest dispersion in Spanish is found in the deontic markers, whereas epistemic ones are most concentrated ones.

Table 59: Column statistics of epistemic vs deontic markers in Spanish and Japanese corpora

	Spanish			Japanese		
Stat	**Epistemic**	**Deontic**	**Ambg**	**Epistemic**	**Deontic**	**Ambg**
N	379	379	379	58	58	58
Min.	0.0	0.0	0.0	0.0	0.0	0.0
25% Percn	0.0	0.0	0.0	0.8216	1.569	0.0
Median	0.0	4.065	2.596	2.986	3.741	0.0
75% Percn	2.229	9.270	5.825	5.333	5.788	0.0
Max.	26.09	85.94	36.04	11.71	13.74	0.0
Mean	1.729	6.365	3.828	3.470	4.141	0.0
SD	3.302	8.786	4.864	3.103	3.219	0.0

Figure 18: Dispersion of epistemic, deontic and ambiguous markers in Spanish and Japanese corpora

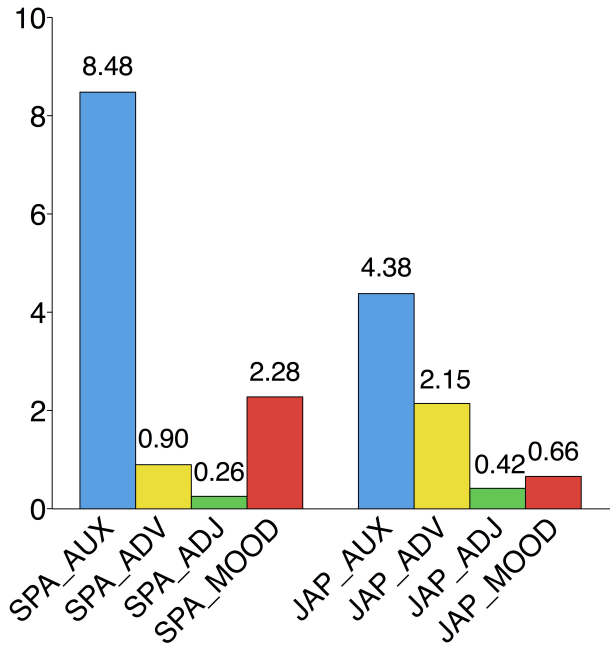


4.2.4 Type of modal markers

Finally we will look at the different types of modal markers used in each corpora. The overall means show that auxiliaries are the most frequent elements to express modality in Spanish and Japanese (see Figure 19), especially in Spanish with a mean of 8.48 usages per speaker per 1000 words, compared to the 4.38 in Japanese.

The second most used type of marker in Spanish is mood, either the imperative, or the negative subjunctive. Their number however is much lower than the auxiliaries, with a mean of 2.28 (nearly 4 times lower). In Japanese the second marker is the adverb, with a figure closer to the auxiliaries (2.15). In Spanish adverbs are not used so frequently, achieving the third place in terms of frequency, with a mean of 0.90, in the same way as mood in Japanese. Finally, the least used modal marker type in both languages is the predicative adjective, although it is slightly higher in Japanese.

Figure 19: Frequency of markers according to their grammatical type



The difference from the number of auxiliary markers and the second most used type of marker is significant in both languages, confirming the fourth hypothesis (Table 60).

The dispersion seems to be very similar to the previous cases, with much more

Table 60: t tests results for Spanish and Japanese auxiliaries and the next most frequent marker

Comparison	**Feature**	**Result**
Auxiliaries - Mood (SPA)	P value	< 0.0001
	P summary	***
	Signf. different	Yes
Auxiliaries - Adverbs (JAP)	P value	< 0.0001
	P summary	***
	Signf. different	Yes

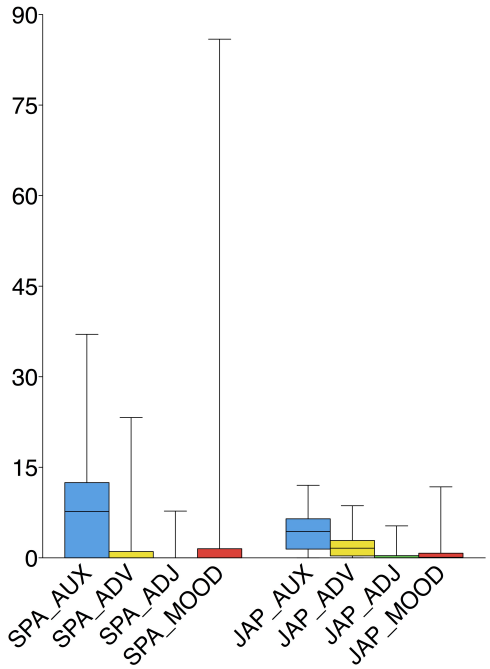
concentrated and regular values in Japanese than in Spanish (see Table 61 and Figure 20). In this case we can observe that the abnormal maximum values that appeared in the total modality numbers in Spanish belong mainly to mood (we may recall that it is due to speaker 'MAM_SPAwoman_C_Segovia', which uses 10 imperatives in a speech with only 128 words). The Spanish mood figures however, aside from the atypical maximum value, seem very concentrated, with a median of 0, a mean of 2.3 and the 75% of the cases situated around the 1.5 usages per speaker.

This is not the case with Spanish auxiliaries, which have a high maximum value of 37 but also a relatively wide dispersion in comparison to the rest of markers. It has a median of 7.7 uses per speaker, a SD of 7.8 and a 75% percentile value of 12.57, compared to the 0.0 value of the 25%. Japanese auxiliaries are also the less concentrated marker in that language, with a median of 2.83 and a SD of 2.77. Overall it appears that the Dispersion of modality of auxiliaries is more diverse between the speakers.

Table 61: Column statistics of the type of modal markers in Spanish and Japanese corpora

	Spanish				Japanese			
Stat	**AUX**	**ADV**	**ADJ**	**MOOD**	**AUX**	**ADV**	**ADJ**	**MOOD**
N	379	379	379	379	58	58	58	58
Min.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25% Percn.	0.0	0.0	0.0	0.0	1.467	0.3253	0.0	0.0
Median	7.692	0.0	0.0	0.0	4.364	1.598	0.0	0.1063
75% Percn	12.47	1.055	0.0	1.513	6.485	2.874	0.3822	0.7689
Max.	37.04	23.26	7.752	85.94	12.03	8.642	5.305	11.76
Mean	8.483	0.9011	0.2578	2.281	4.383	2.147	0.4188	0.6635
SD	7.888	2.186	0.8731	7.495	3.212	2.291	0.9938	1.629

Figure 20: Frequency dispersion of the type of markers used in Spanish and Japanese corpora



4.3 Modality frequency: linguistic factors

This section will describe the frequencies of modal markers according to the type of discourse (monologues and dialogues in Spanish and Japanese), and register differences in formal and informal Spanish texts. Table 62 presents the distribution of the file types among both corpora. Considering the low amount of formal texts in Japanese and the lack of media texts, the cross-linguistic comparisons will be made strictly between dialogues and monologues. Subsection 4.3.2 will briefly present the frequencies among informal and formal Spanish texts.

Table 62: Breakdown of files and words of the corpora

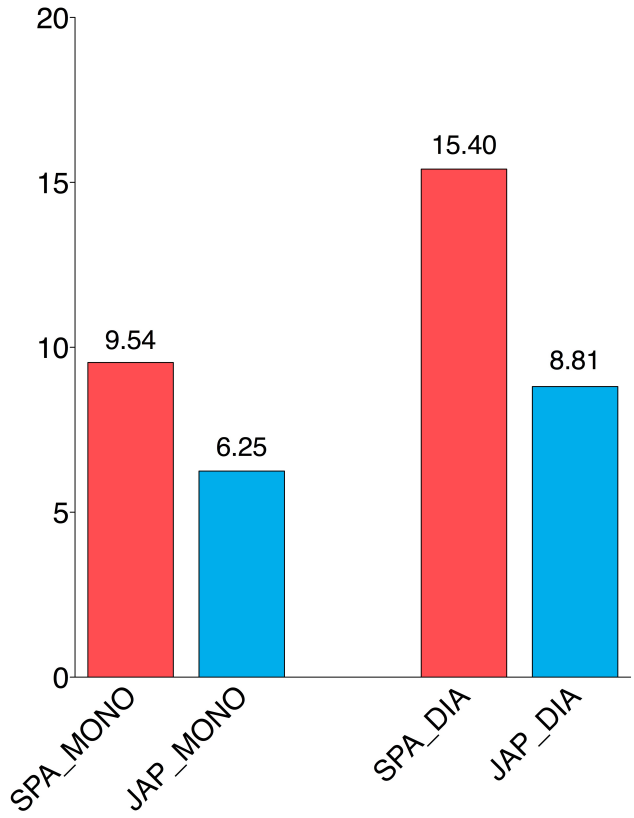
		Informal		**Formal**		
Corpus	**Files**	**Monolgs**	**Dialgs**	**Monolgs**	**Dialgs**	**Media**
CORAL-ROM	169	12(47,138)	87(128,723)	17(36,938)	11(28,729)	42(59,801)
CORAL-JP	39	11(48,912)	24(63,994)	4(14,770)	0	0

4.3.1 Discourse type

4.3.1.1 Modality in monologues and dialogues

First of all, Figure 21 indicates a higher usage of modality in dialogues than in monologues. The monologue-dialogue difference in both languages is quite similar (1.6 in Spanish, 1.4 in Japanese). In dialogues, Spanish speakers appear to use around 15 instances of modality per 1000 words, compared to Japanese that use nearly 9 markers per speaker. The usage in monologues is closer, with an average of three more markers in Spanish than in Japanese.

Figure 21: Markers per speaker (mean) in Spanish and Japanese monologues and dialogues



The t tests (Table indicate 63) show that the difference between the monologue and dialogue means is significant in both languages, especially Spanish.

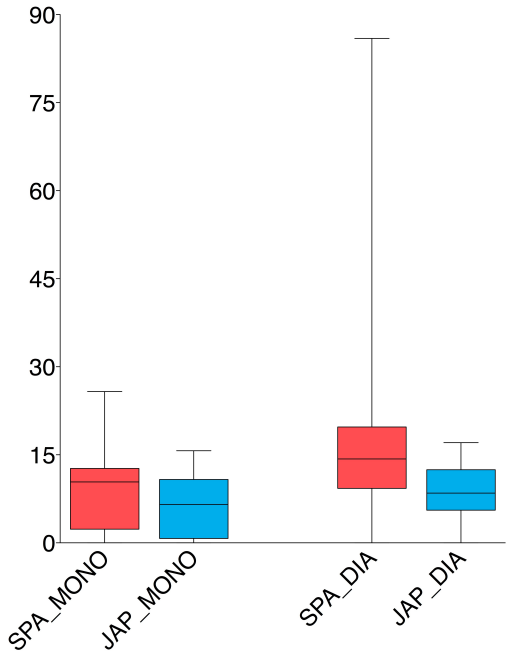
Table 63: t tests (two-tailed) results for modality according to the type of discourse

Comparison	Feature	Result
Monologues - Dialogues (SPA)	P value	0.0008
	P summary	***
	Signf. different	Yes
Monologues - Dialogues (JAP)	P value	0.0392
	P summary	*
	Signf. different	Yes

Observing the dispersion, we see similar results: Spanish frequencies are less concentrated than their Japanese counterparts, as displayed below in Figure 22 and Table 64, compared to the general figures seen before (Figure 12 and Table 54).

Japanese frequencies pass again the three normality tests, and there is again an atypical high value in Spanish dialogues.

Figure 22: Dispersion of modality in Spanish and Japanese monologues and dialogues



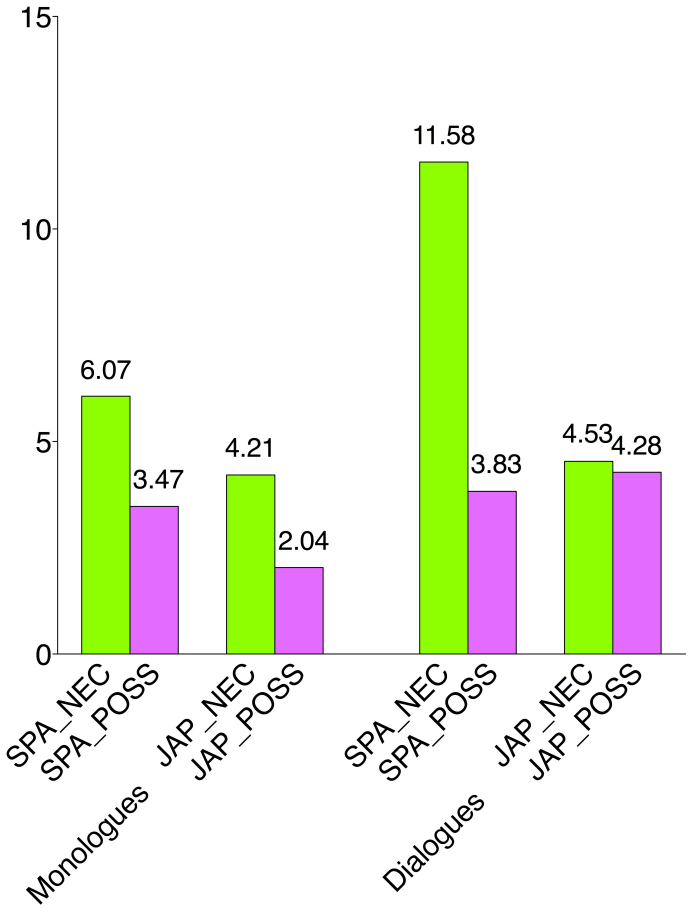
Nevertheless, the differences are not so high. Spanish monologues have passed one of the normality tests, and have achieved higher P values in the other two (0.01 and 0.003). We still cannot consider them as balanced as Japanese, but they are more evenly distributed. In terms of deviation, Spanish dialogues have the highest values and have an abnormal maximum of 86 markers. In the rest of situations the maximum points achieve similar situations and do not surpass the 25 markers. Spanish and Japanese monologues have SDs of 5 and 7, 75% percentile of 13 and 11 markers respectively. The majority of Spanish monologues are situated between the 25% and the median, although Japanese data is more symmetric. Also, Spanish data fails to pass normality tests again, but monologues have higher P values.

Table 64: Column statistics of modality in Spanish and Japanese monologues and dialogues

	Monologue		**Dialogue**	
Stat	**Spanish**	**Japanese**	**Spanish**	**Japanese**
N	37	25	182	34
Min.	0.0	0.0	0.0	0.0
25% Percn.	2.329	0.7576	9.278	5.563
Median	10.38	6.557	14.31	8.480
75% Percn	12.68	10.79	19.73	12.46
Max.	25.77	15.70	85.94	17.06
Mean	9.540	6.251	15.40	8.810
SD	7.050	4.970	10.93	4.313
KS test				
P value	0.0246	> 0.10	< 0.0001	> 0.10
Passed?	No	Yes	No	Yes
P value	*	ns	***	ns
D'Agostino and P. test				
P value	0.6614	0.3015	< 0.0001	0.5942
Passed?	Yes	Yes	No	Yes
P value	ns	ns	***	ns
Shapiro-Wilk test				
P value	0.0038	0.0658	< 0.0001	0.8398
Passed?	No	Yes	No	Yes
P value	**	ns	***	ns

When comparing the average necessity and possibility markers per speaker, necessity is higher in every case. The difference of the means is similar in monologues, but completely different in dialogues, where necessity markers are four times higher in Spanish dialogues but nearly equal to possibility in Japanese (Figure 23).

Figure 23: Necessity/Possibility mean per speaker in Spanish and Japanese monologues and dialogues

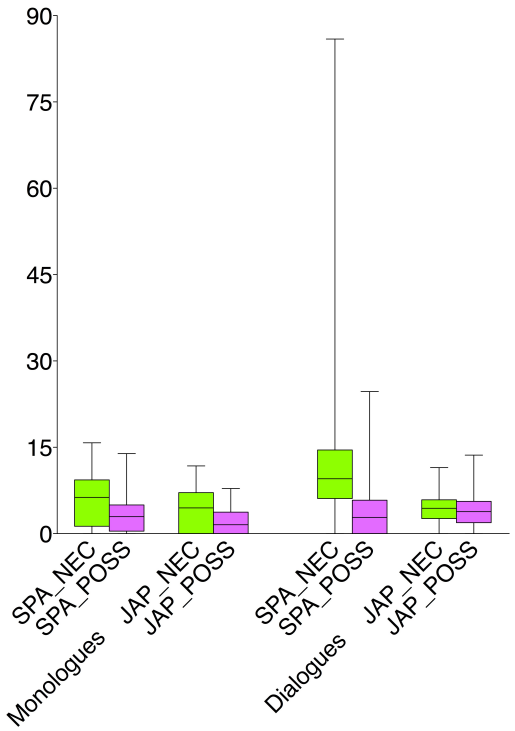


Regarding their dispersion, every type of discourse achieves similar results, with the exception of necessity markers in Spanish dialogues (see Table 65 and Figure 23). Possibility markers are in average more concentrated in every situation. The most remarkable feature (aside from the higher maximums found in Spanish dialogues) is the higher values of possibility markers in Japanese dialogues: the 25% percentile is situated in nearly 2 markers per speaker, a median of 3.8 and a mean of 4.3 markers, higher than the rest of discourses.

Table 65: Column statistics of necessity/possibility in Spanish and Japanese monologues and dialogues

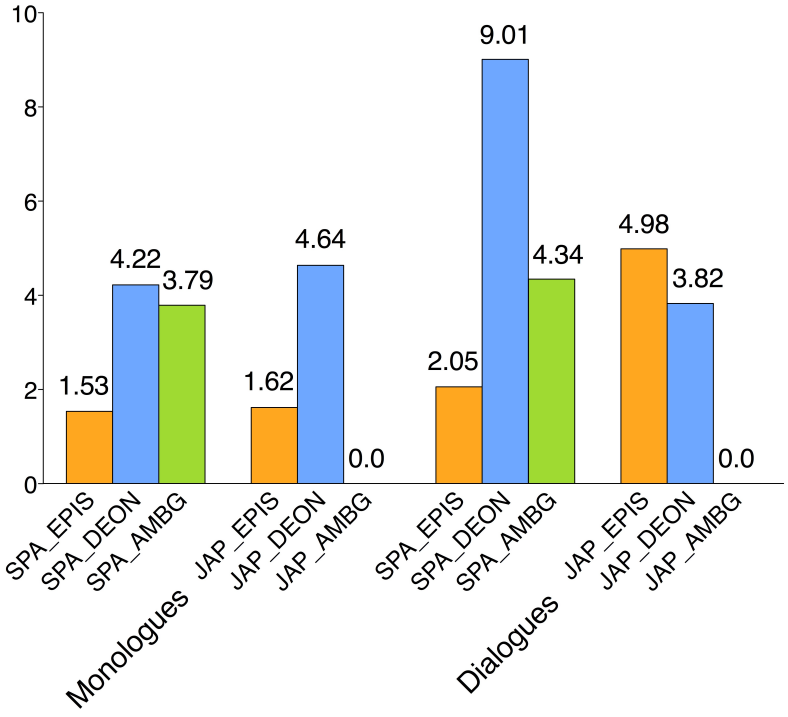
	Monologues				**Dialogues**			
	Spanish		**Japanese**		**Spanish**		**Japanese**	
Stat	**NEC**	**POSS**	**NEC**	**POSS**	**NEC**	**POSS**	**NEC**	**POSS**
N	37	37	25	25	182	182	34	34
Min.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25% Percn.	1.284	0.4368	0.0	0.0	6.097	0.0	2.615	1.912
Median	6.310	2.955	4.480	1.553	9.532	2.813	4.410	3.820
75% Percn	9.344	4.977	7.124	3.730	14.53	5.803	5.889	5.610
Max.	15.78	13.92	11.76	7.851	85.94	24.69	11.48	13.65
Mean	6.067	3.473	4.214	2.037	11.58	3.828	4.534	4.276
SD	4.536	3.461	3.623	2.171	10.45	4.236	2.577	3.237

Figure 24: Dispersion of necessity/possibility frequency in Spanish and Japanese monologues and dialogues



Looking at the frequencies of epistemic and deontic modalities, the numbers remain very similar to the general ones seen in Figure 11. Overall, deontic is higher than epistemic modality and the difference is quite noticeable, especially in Spanish dialogues. There is an exception, however, in Japanese dialogues, where epistemic modality is an average of 1 marker higher than deontic modality. Speakers in Japanese have a mean of nearly 5 epistemic markers per speaker, much higher than the rest of cases (nearly four times higher than monologues and double the amount of Spanish dialogues). It seems deontic markers are used more freely when speaking alone (either the addressee of the modality is the speaker, or he/she is speaking in third person) than in a dialogue. Therefore, our third hypothesis, which stated that deontic markers would be lower in Japanese, is not entirely rejected.

Figure 25: Epistemic/Deontic frequency in Spanish and Japanese monologues and dialogues



Continuing with these assumptions, in terms of dispersion, represented in Tables 66 and 67 and Figure 26, Japanese dialogues are again the exception: in these discourses deontic modality is the most concentrated, compared to the other cases where the epistemic numbers are more concentrated than deontic. Monologue figures are quite similar in both languages, but they differ mostly in the dialogues.

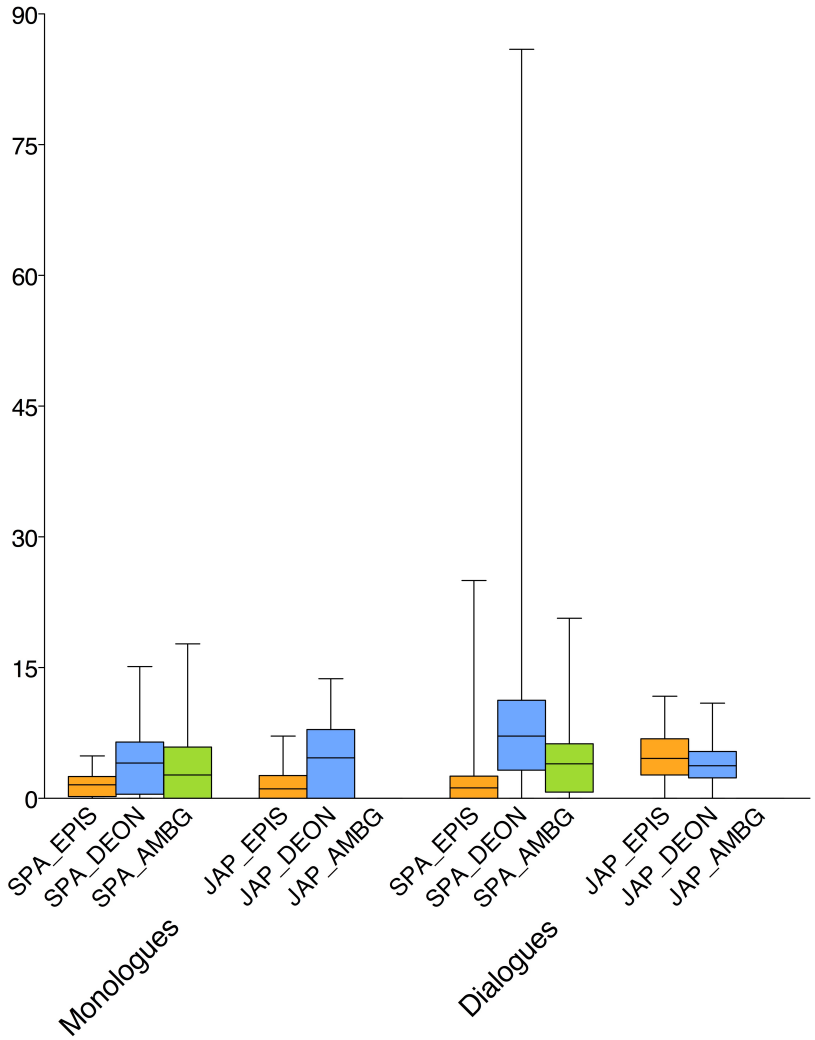
Table 66: Column statistics of epistemic/deontic modality in Spanish and Japanese monologues

	Spanish			**Japanese**		
Stat	**EPIS**	**DEON**	**AMBG**	**EPIS**	**DEON**	**AMBG**
N	37	37	37	25	25	25
Min.	0.0	0.0	0.0	0.0	0.0	0.0
25% Percn.	0.2184	0.4794	0.0	0.0	0.0	0.0
Median	1.552	4.044	2.673	1.093	4.658	0.0
75% Percn	2.498	6.458	5.879	2.621	7.887	0.0
Max.	4.866	15.12	17.72	7.143	13.74	0.0
Mean	1.533	4.220	3.787	1.616	4.636	0.0
SD	1.332	3.705	4.160	1.923	4.197	0.0

Table 67: Column statistics of epistemic/deontic modality in Spanish and Japanese dialogues

	Spanish			**Japanese**		
Stat	**EPIS**	**DEON**	**AMBG**	**EPIS**	**DEON**	**AMBG**
N	182	182	182	34	34	34
Min.	0.0	0.0	0.0	0.0	0.0	0.0
25% Percn.	0.0	3.232	0.7156	2.675	2.347	0.0
Median	1.198	7.145	3.957	4.589	3.741	0.0
75% Percn	2.551	11.25	6.273	6.837	5.385	0.0
Max.	25.00	85.94	20.68	11.71	10.93	0.0
Mean	2.053	9.008	4.342	4.983	3.827	0.0
SD	3.088	10.18	3.995	3.147	2.259	0.0

Figure 26: Epistemic/Deontic dispersion in Spanish and Japanese monologues and dialogues



Finishing the section with the type of markers, monologues are very similar in both languages: auxiliaries are the most frequent marker, followed by adverbs and mood. Dialogues, on the other hand show another picture. Auxiliaries receive the highest numbers, but mood is much more frequent in Spanish than in Japanese, where the second most frequent class of marker is the adverb. On average, modal adjectives are more used than mood in Japanese (Figures 27 and 28).

Figure 27: Mean per speaker of modal marker type in Spanish and Japanese monologues

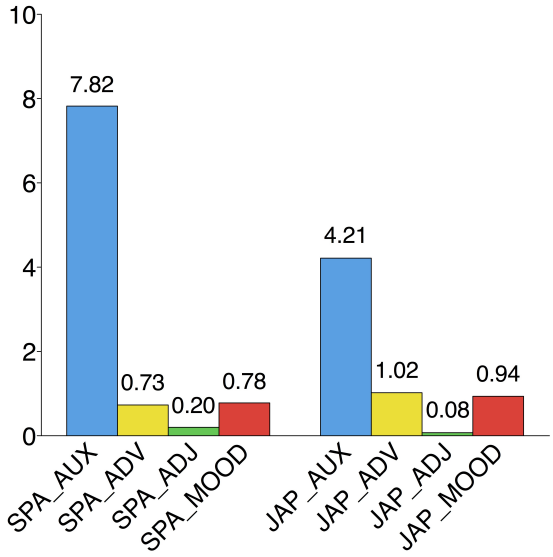
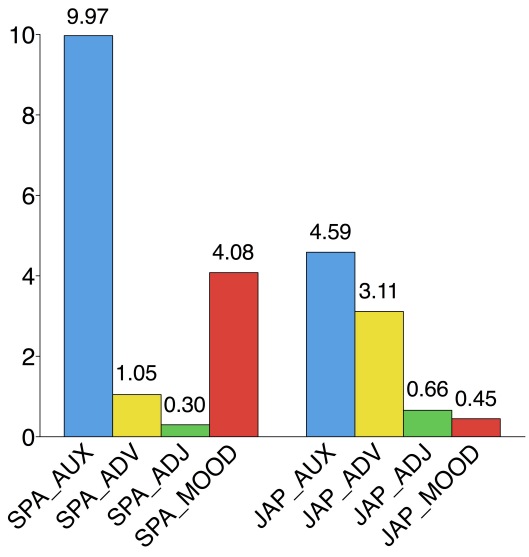


Figure 28: Mean per speaker of modal marker type in Spanish and Japanese dialogues



The dispersion shows similar results in all cases: the most noticeable difference can be found in the dialogues (Tables 68 and 69; Figures 29 and 30). Japanese are more regular and concentrated, but Spanish are more diverse, especially mood, with the abnormal maximum.

Table 68: Column statistics of modal markers in Spanish and Japanese monologues

	Spanish				**Japanese**			
Stat	**Aux**	**Adv.**	**Adj.**	**Mood**	**Aux.**	**Adv.**	**Adj.**	**Mood**
N	37	37	37	37	25	25	25	25
Min.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25% P.	1.663	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Median	8.303	0.4396	0.0	0.0	3.690	0.4422	0.0	0.2125
75% P	11.12	1.325	0.3178	0.9990	8.211	1.372	0.0	1.038
Max.	23.52	3.650	1.898	6.575	11.78	7.143	0.4907	11.76
Mean	7.823	0.7314	0.2043	0.7812	4.214	1.022	0.07786	0.9373
SD	5.846	0.8489	0.4356	1.422	3.896	1.629	0.1620	2.348

Table 69: Column statistics of modal markers in Spanish and Japanese dialogues

	Spanish				**Japanese**			
Stat	**Aux**	**Adv.**	**Adj.**	**Mood**	**Aux.**	**Adv.**	**Adj.**	**Mood**
N	182	182	182	182	34	34	34	34
Min.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25% P.	4.970	0.0	0.0	0.0	2.935	1.487	0.0	0.0
Median	9.509	0.0	0.0	1.146	4.514	2.690	0.0	0.0
75% P	14.06	1.520	0.0	3.318	6.403	3.765	0.7070	0.6899
Max.	34.48	12.35	4.032	85.94	12.03	8.642	5.305	3.413
Mean	9.971	1.048	0.3031	4.081	4.589	3.117	0.6584	0.4465
SD	6.971	1.792	0.7279	10.20	2.652	2.454	1.243	0.6942

Figure 29: Dispersion of modal markers in monologues

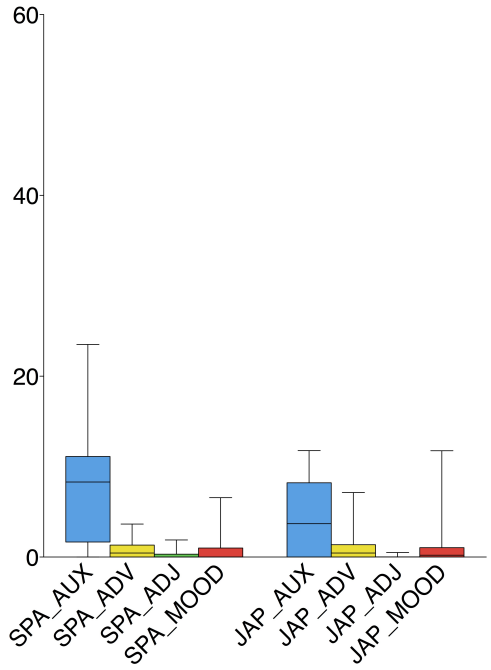
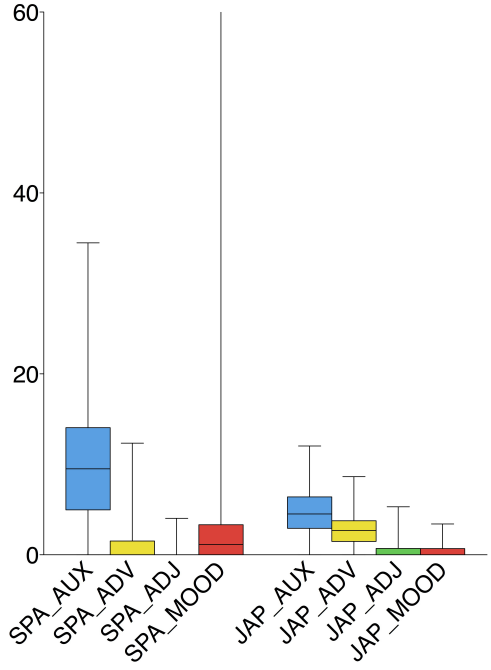


Figure 30: Dispersion of modal markers in dialogues



4.3.2 Register

Moving now to the type of register of the texts, whether they are formal or informal monologues or dialogues, the Japanese numbers are too low to perform an appropriate statistical study, which has forced us to consider only Spanish for the study.

Overall, there is a higher amount of modality usage in informal (mean of 15.05) than formal Spanish (9.8) (see Figure 31). The difference is only an average of 5 markers, but the t test indicates it is a significant difference of usage.

Table 70: Mann Whitney test (two-tailed) results for formal and informal Spanish

Feature	Result
P value	< 0.0001
P summary	***
Signf. different	Yes

Figure 31: Mean per speaker of modality in informal vs formal spoken Spanish

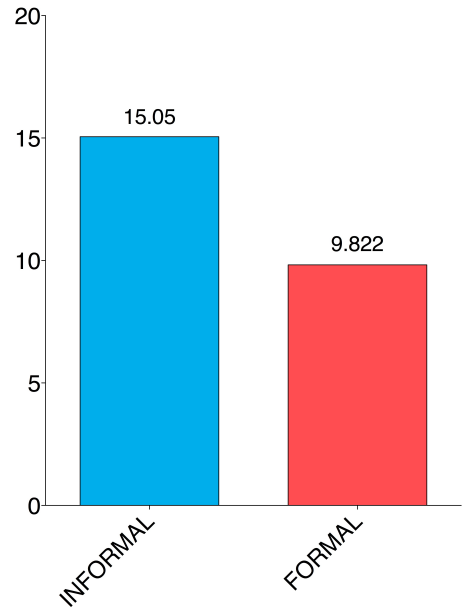
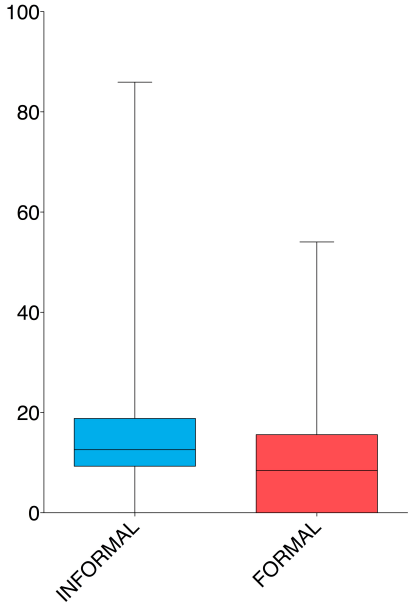


Figure 32: Modality dispersion in informal vs formal spoken Spanish



In terms of dispersion (Figure 32 above and Table 71 below), informal Spanish frequencies are also more concentrated than the formal ones, with the majority of users using an average of 9 to 19 modality instances per 1000 words, whereas the majority range of formal Spanish is more diverse, from 0 to around 19 cases. Both type of texts have high maximum points, and neither pass the normal distribution test.

The necessity and possibility values show hardly any surprises (Figures 33 and 34 and Table 72), with higher frequencies of necessity than possibility in both registers, especially in the informal register. The difference from possibility is smaller in the case of formal interactions (3 times higher in informal cases vs 1.5 times in formal ones).

Table 71: Column statistics of modality in informal vs formal spoken Spanish

Stat	**Informal**	**Formal**
N	167	222
Minimum	0.0	0.0
25% P.	9.288	0.0
Median	12.95	7.830
75% P.	18.85	15.53
Maximum	85.94	54.05
Mean	15.03	9.578
Std. Deviation	10.88	9.614
KS normality test		
P value	< 0.0001	< 0.0001
Passed normality test (alpha=0.05)?	No	No
P value summary	***	***
D'Agostino and Pearson test		
P value	< 0.0001	< 0.0001
Passed normality test (alpha=0.05)?	No	No
P value summary	***	***
Shapiro-Wilk normality test		
P value	< 0.0001	< 0.0001
Passed normality test (alpha=0.05)?	No	No
P value summary	***	***

Figure 33: Mean per speaker of necessity and possibility markers in informal vs formal Spanish

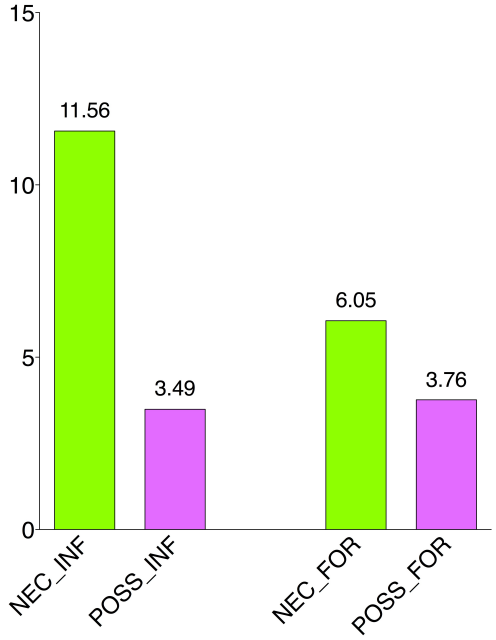


Figure 34: Dispersion of necessity and possibility markers in informal vs formal Spanish

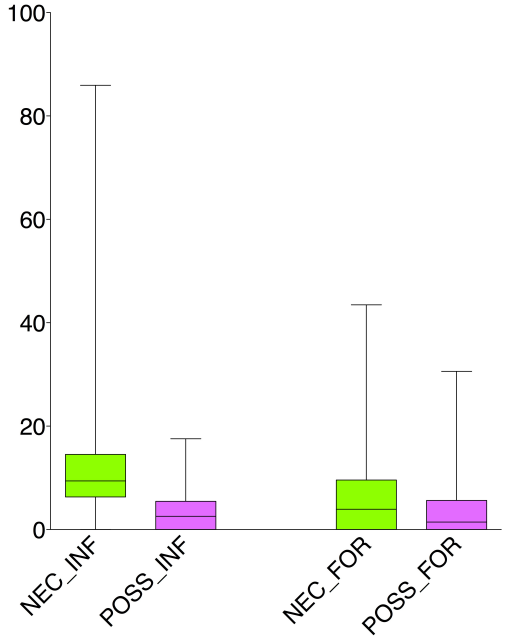


Table 72: Column statistics of necessity vs possibility in informal vs formal Spanish

	Informal Spanish		**Formal Spanish**	
Stat	**NEC**	**POSS**	**NEC**	**POSS**
N	162	162	224	224
Min.	0.0	0.0	0.0	0.0
25% Percn.	6.319	0.0	0.0	0.0
Median	9.396	2.569	3.917	1.452
75% Percn	14.53	5.455	9.576	5.630
Max.	85.94	17.57	43.48	30.61
Mean	11.56	3.490	6.058	3.764
SD	10.44	3.805	7.687	5.603

The amount of deontic markers appears to be directly related to those of necessity: their number is considerably higher in informal situations than any other

type of modality (Figure 35), and their dispersion is relatively small (Table 73 and Figure 36). Ambiguity is also high and nearly the same amount in both type of registers (Figure 35), though more dispersed in formal conversations (Figure 36 and Table 73).

Figure 35: Mean per speaker of epistemic, deontic and ambiguous markers in informal vs formal Spanish

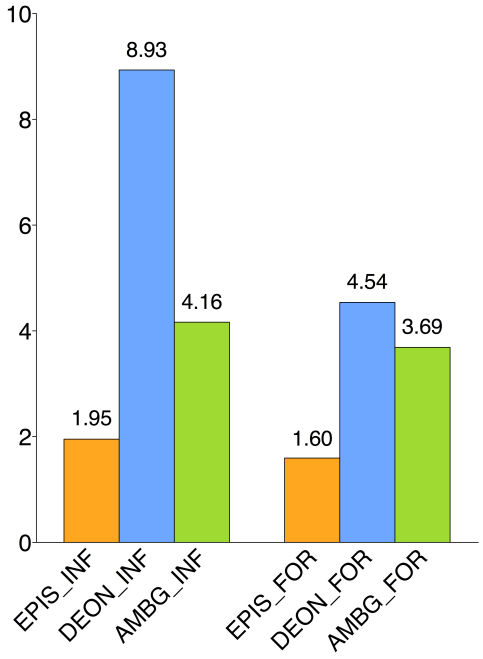


Figure 36: Dispersion of epistemic, deontic and ambiguous markers in informal vs formal Spanish

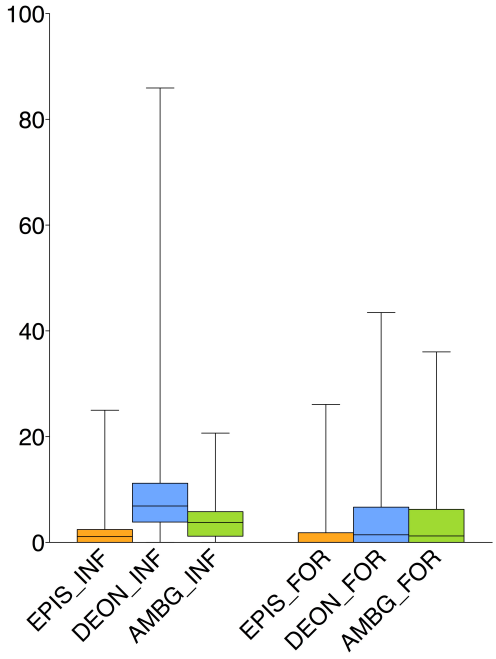


Table 73: Column statistics of epistemic vs deontic markers in informal vs formal Spanish

	Informal			**Formal**		
Stat	**Epistemic**	**Deontic**	**Ambiguous**	**Epistemic**	**Deontic**	**Ambiguous**
N	162	162	162	224	224	224
Min.	0.0	0.0	0.0	0.0	0.0	0.0
25% Percn.	0.0	3.856	1.178	0.0	0.0	0.0
Median	1.118	6.879	3.773	0.0	1.448	1.218
75% Percn	2.453	11.20	5.829	1.833	6.684	6.252
Max.	25.00	85.94	20.68	26.09	43.48	36.04
Mean	1.954	8.934	4.165	1.595	4.539	3.688
SD	3.005	10.10	3.711	3.529	7.121	5.572

The same conclusions can be drawn from the grammatical category of the markers: auxiliaries are much more frequent than any other kind of modal marker, and slightly more used in informal than formal situations (Figure 37). The difference is not very high, an average of 9.70 in the former vs a 7.67 in the latter, but the dispersion is higher in the formal situations (Table 74 and Figure 38), very related to the numbers of necessity seen before. It is followed by imperative and negative subjunctive moods, adverbs, and predicative adjectives.

Figure 37: Type of modal markers in informal vs formal spoken Spanish

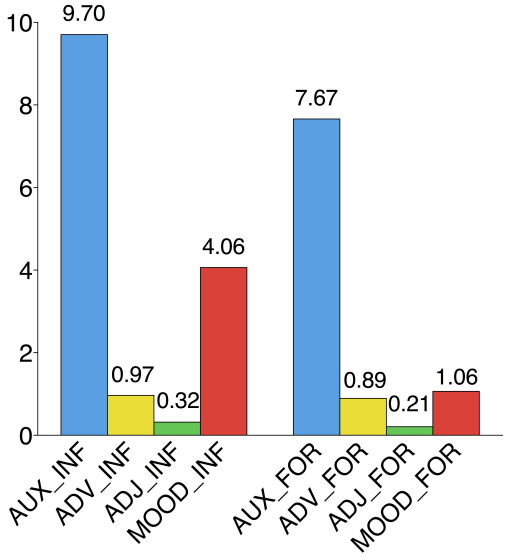


Figure 38: Dispersion of the type of modal markers used in informal vs formal spoken Spanish

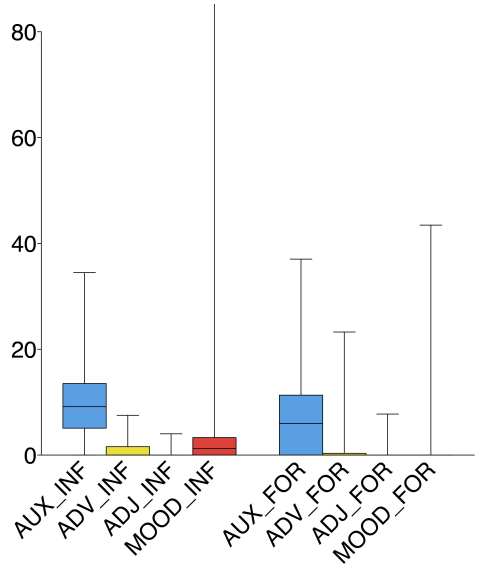


Table 74: Column statistics of modal markers used in informal vs formal spoken Spanish

	Informal Spanish				**Formal Spanish**			
Stat	**AUX**	**ADV**	**ADJ**	**MOOD**	**AUX**	**ADV**	**ADJ**	**MOOD**
N	162	162	162	162	224	224	224	224
Min.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25% Percn.	5.082	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Median	9.158	0.0	0.0	1.252	5.994	0.0	0.0	0.0
75% Percn	13.53	1.594	0.0	3.318	11.33	0.3327	0.0	0.0
Max.	34.48	7.533	4.032	85.94	37.04	23.26	7.752	43.48
Mean	9.704	0.9660	0.3200	4.064	7.661	0.8920	0.2061	1.063
SD	6.823	1.496	0.7573	10.22	8.432	2.662	0.9355	4.264

4.4 Non-linguistic factors

We will now turn to the observation of similarities and differences in modality according to non-linguistic elements such as gender and age of the speakers. Table 75 presents the amount of speakers and their ages (A: 18-25, B: 26-40, C: 41-50, D: 51+). Since the number of speakers per calculation decreases, especially in group D, and there is a majority of Spanish speakers with unknown age, we cannot draw very strong conclusions, and the section will focus mainly on comparing the total counts of the averages.

Table 75: Breakdown of number of speakers per age group

	Women					**Men**				
Corpus	**A**	**B**	**C**	**D**	**Unkn**	**A**	**B**	**C**	**D**	**Unkn**
CORAL-ROM	36	50	28	5	35	24	45	63	3	90
CORAL-JP	20	2	5	10	0	3	12	5	1	0

4.4.1 Modality usage according to gender

Overall, it appears that women use a higher average of modal markers than men, as represented in Figure 39. Nevertheless, the differences are not too deep and quite similar, Spanish women count is 1.11 times higher than men whereas in Japanese it is only 1.15. The t tests, additionally, show these differences are not significant (Table 76).

Both Spanish men and women show less concentration than the Japanese, which once again pass the normality tests. Spanish women frequencies appear more concentrated than men, with averages between 6.13 and 17.41 per speaker compared to the men's averages of 2.84 and 16.78 (Figure 40 and Table 77).

Figure 39: Mean per speaker of modality women and men

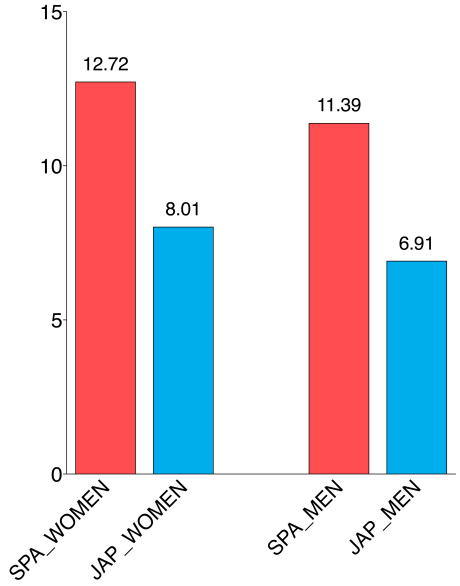


Figure 40: Modality dispersion in women and men

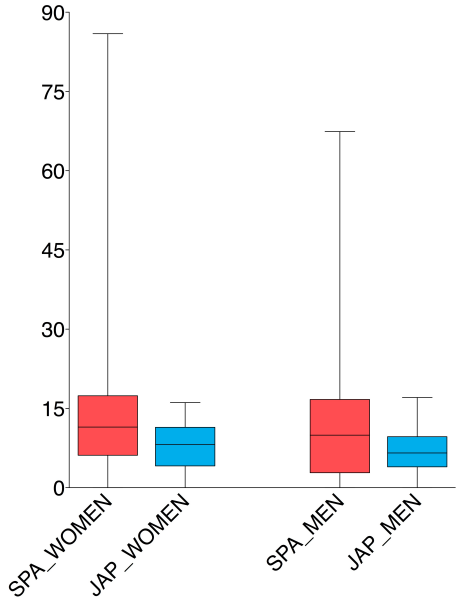


Table 76: t tests (two-tailed) results for modality between men and women in both languages

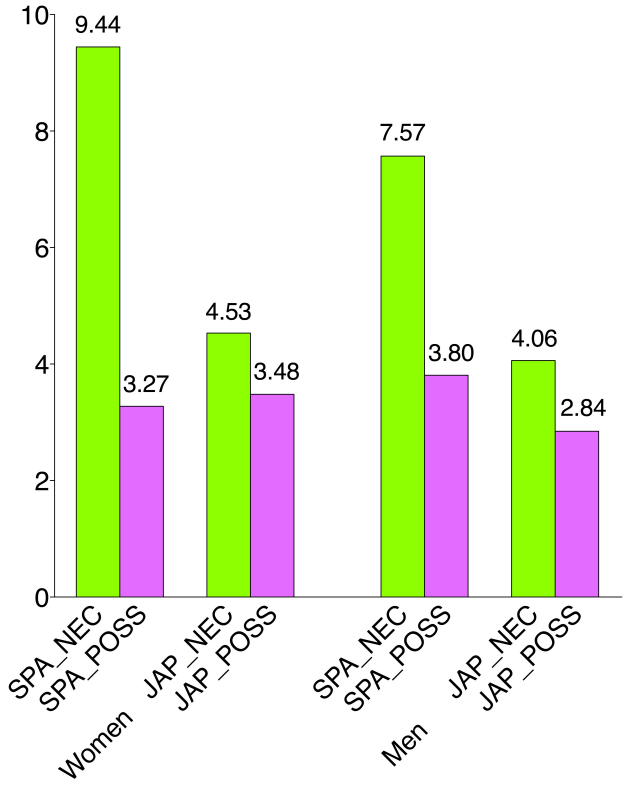
Comparison	**Feature**	**Result**
Men - Women (SPA)	P value	0.1553
	P summary	ns
	Signf. different	No
Men - Women (JAP)	P value	0.3941
	P summary	ns
	Signf. different	No

Table 77: Column statistics of modality in Spanish and Japanese women and men

	Women		**Men**	
Stat	**Spanish**	**Japanese**	**Spanish**	**Japanese**
N	154	37	225	21
Min.	0.0	0.0	0.0	0.0
25% Percn.	6.132	4.116	2.842	3.941
Median	11.47	8.177	9.938	6.579
75% Percn	17.41	11.42	16.71	9.656
Max.	85.94	16.11	67.42	17.06
Mean	12.72	8.012	11.38	6.907
SD	10.99	4.692	10.44	4.730
KS test				
P value	< 0.0001	> 0.10	< 0.0001	> 0.10
Passed?	No	Yes	No	Yes
P value summary	***	ns	***	ns
D'Agostino and Pearson test				
P value	< 0.0001	0.2604	< 0.0001	0.6374
Passed?	No	Yes	No	Yes
P value summary	***	ns	***	ns
Shapiro-Wilk test				
P value	< 0.0001	0.2568	< 0.0001	0.3408
Passed?	No	Yes	No	Yes
P value summary	***	ns	***	ns

We can observe a similar pattern in the necessity vs possibility modality comparison (Figure 41 and Table 78). Both Spanish and Japanese women seem to use more necessity markers than men, and less possibility markers than them, although the difference is lower in Japanese. Also, the possibility values are closer in women, slightly higher, for the first time, in Japanese.

Figure 41: Necessity/Possibility frequency in Spanish and Japanese women and men



Regarding the division between epistemic and deontic modality (Figure 42 and Tables 79 and 80), in all cases the most used type is deontic modality. However, the difference in usage depends on the situation: Spanish women and men use similar amounts of epistemic and ambiguous markers. The difference from deontics is very wide, specifically in women. Japanese women and men, on the other hand, are similar in deontic markers, and women use more frequently epistemic ones than men. The epistemic-deontic dissimilarity is smaller, especially among women, who use nearly the same amount of markers.

Table 78: Column statistics of necessity/possibility in Spanish and Japanese women and men

	Monologues				**Dialogues**			
	Spanish		**Japanese**		**Spanish**		**Japanese**	
Stat	**NEC**	**POSS**	**NEC**	**POSS**	**NEC**	**POSS**	**NEC**	**POSS**
N	154	154	37	37	225	225	21	21
Min.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25% Percn.	3.307	0.0	2.269	1.190	0.0	0.0	1.754	0.0
Median	8.202	1.751	4.375	3.511	6.098	2.248	4.480	1.618
75% Percn	13.18	5.267	6.331	4.781	10.98	5.722	6.292	3.962
Max.	85.94	27.03	11.48	10.25	67.42	30.61	11.76	13.65
Mean	9.443	3.275	4.531	3.481	7.571	3.807	4.061	2.846
SD	9.515	4.435	3.022	2.742	9.205	5.094	3.101	3.397

Figure 42: Epistemic/Deontic frequency in Spanish and Japanese women and men

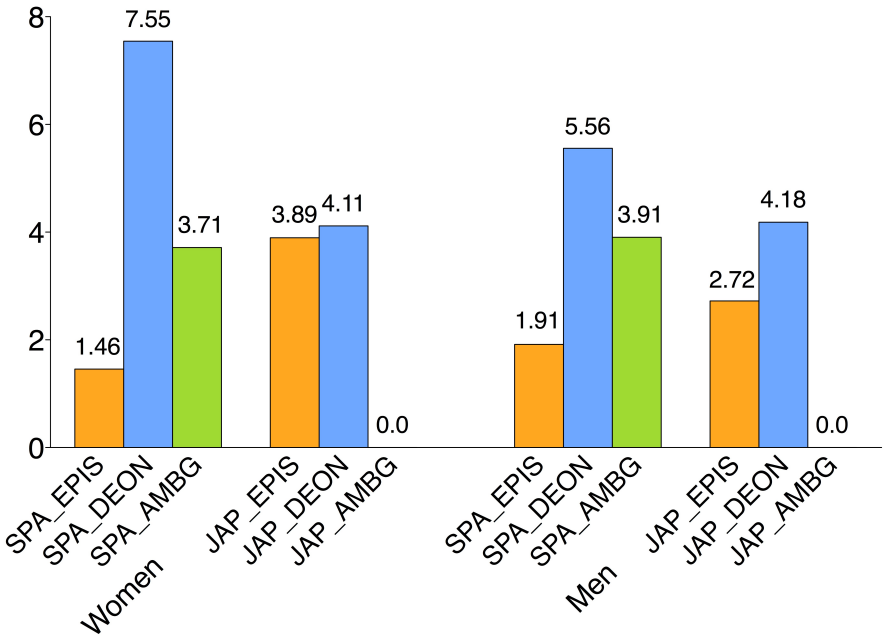


Table 79: Column statistics of epistemic/deontic modality in Spanish and Japanese women

	Spanish			Japanese		
Stat	**EPIS**	**DEON**	**AMBG**	**EPIS**	**DEON**	**AMBG**
N	154	154	154	37	37	37
Min.	0.0	0.0	0.0	0.0	0.0	0.0
25% Percn.	0.0	1.165	0.0	1.169	1.454	0.0
Median	0.0	6.108	2.808	3.849	3.748	0.0
75% Percn	1.806	10.91	5.311	5.534	6.145	0.0
Max.	23.26	85.94	36.04	11.71	10.93	0.0
Mean	1.456	7.547	3.715	3.895	4.117	0.0
SD	2.763	9.104	4.646	3.004	3.051	0.0

Table 80: Column statistics of epistemic/deontic modality in Spanish and Japanese men

	Spanish			Japanese		
Stat	**EPIS**	**DEON**	**AMBG**	**EPIS**	**DEON**	**AMBG**
N	225	225	225	21	21	21
Min.	0.0	0.0	0.0	0.0	0.0	0.0
25% Percn.	0.0	0.0	0.0	0.0	1.980	0.0
Median	0.0	3.080	2.484	1.963	3.604	0.0
75% Percn	2.507	7.294	6.290	3.542	5.613	0.0
Max.	26.09	67.42	26.32	10.24	13.74	0.0
Mean	1.915	5.557	3.906	2.723	4.184	0.0
SD	3.618	8.488	5.017	3.204	3.573	0.0

The grammatical class of markers show the same pattern observed until now: the auxiliaries are the most frequent marker type followed by mood in Spanish and adverbs in Japanese (Figures 43 and 44, Tables 81 and 82). Spanish women and men's figures are very similar, but not in the case of Japanese. Here, the usage is different in both sexes: auxiliaries are the most frequent followed by adverbs, but mood markers have a mean of 1.11 per male speaker, whereas in women it is 0.41, lower than adjectives. There is a slight difference in Japanese women markers, as mood markers are the least frequent ones, below adjectives. Also, both Spanish women and men have a high abnormal maximum point and a fairly dispersed auxiliary numbers. The rest of markers are quite concentrated around their SDs.

Figure 43: Modal markers frequency in Spanish and Japanese women

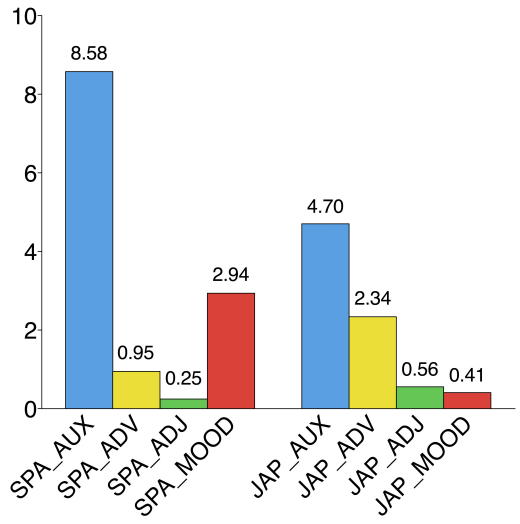


Figure 44: Modal markers frequency in Spanish and Japanese men

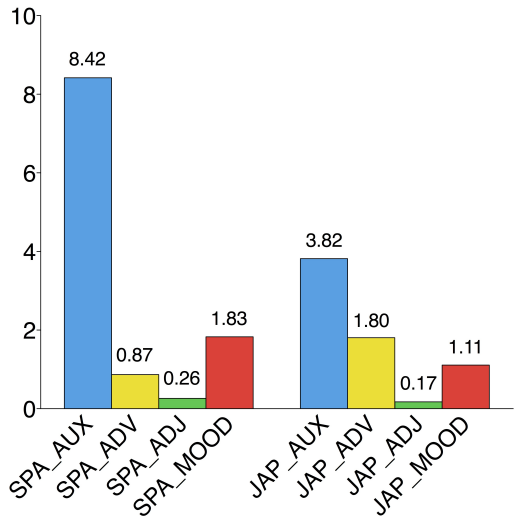


Table 81: Column statistics of modal markers in Spanish and Japanese women

	Spanish				Japanese			
Stat	**Aux**	**Adv.**	**Adj.**	**Mood**	**Aux.**	**Adv.**	**Adj.**	**Mood**
N	154	154	154	154	37	37	37	37
Min.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25% P.	2.750	0.0	0.0	0.0	1.637	0.4856	0.0	0.0
Median	7.930	0.0	0.0	0.0	4.612	2.187	0.0	0.3253
75% P	12.33	1.175	0.0	2.488	6.743	3.440	0.5398	0.7583
Max.	36.04	23.26	3.096	85.94	12.03	8.455	5.305	1.464
Mean	8.577	0.9509	0.2517	2.939	4.704	2.341	0.5573	0.4092
SD	7.575	2.336	0.6435	8.422	3.160	2.291	1.190	0.4717

Table 82: Column statistics of modal markers in Spanish and Japanese men

	Spanish				Japanese			
Stat	**Aux**	**Adv.**	**Adj.**	**Mood**	**Aux.**	**Adv.**	**Adj.**	**Mood**
N	225	225	225	225	21	21	21	21
Min.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25% P.	0.0	0.0	0.0	0.0	0.7576	0.0	0.0	0.0
Median	7.453	0.0	0.0	0.0	3.193	1.327	0.0	0.0
75% P	12.65	0.9574	0.0	0.7601	5.679	2.442	0.2773	1.231
Max.	37.04	17.39	7.752	67.42	11.78	8.642	1.802	11.76
Mean	8.418	0.8670	0.2620	1.831	3.816	1.805	0.1748	1.112
SD	8.112	2.081	1.002	6.772	3.302	2.307	0.4091	2.613

4.4.2 Modality usage according to age

Observing the frequency of modal markers across different ages, we cannot find any noticeable differences of usage among the speakers (See Figure 45). Modality appears to be higher among the youngest Japanese speakers, with an average of 9.35 per speaker per 1000 words. The highest mean among Spanish is found in the oldest speakers (15.23). Nevertheless, the dissimilarity in the frequencies is not very high. Tests (Table 83) comparing the four means in each language show a lack of significance among them.

Figure 45: Modal markers frequency in Spanish and Japanese age groups

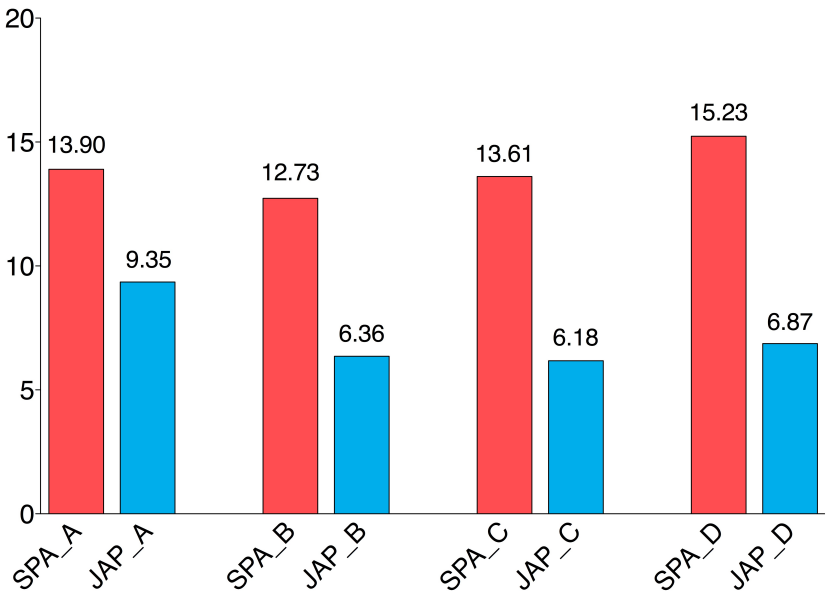


Table 83: ANOVA tests results for modality among 4 age groups in both languages

Comparison	Feature	Result
Ages - Spanish	P value	0.9487
	P summary	ns
	Signf. different	No
Ages - Japanese	P value	0.1452
	P summary	ns
	Signf. different	No

In terms of dispersion, the stand-out feature is the Spanish speakers in the B group passing three normality tests, although not achieving a high significance (p values >0.10, 0.57 and 0.66 respectively). Apart from this, there are hardly any other new features: small dispersion, especially among Japanese speakers, which have passed the normal distribution tests, and high, abnormal maximum points in Spanish speakers, particularly A and C speakers (Table 45).

Table 84: Column statistics of modal markers in Spanish and Japanese age groups

	Spanish				**Japanese**			
Stat	**A**	**B**	**C**	**D**	**A**	**B**	**C**	**D**
N	60	23	95	14	91	10	8	11
Min.	0.0	0.0	0.0	0.0	0.0	0.0	8.717	0.0
25% P.	9.310	6.740	6.390	1.136	6.289	2.427	9.810	3.891
Median	12.41	8.783	11.69	6.860	12.87	5.745	11.14	6.557
75% P.	17.64	13.84	18.54	10.40	17.98	8.423	17.22	10.73
Max.	67.42	17.06	34.78	12.82	85.94	15.70	40.00	11.81
Mean	13.90	9.353	12.73	6.358	13.61	6.178	15.23	6.870
SD	10.05	4.812	8.748	4.481	11.92	4.910	10.49	3.877
KS t.								
P value	0.0031	> 0.10	0.0865	> 0.10	0.0010	> 0.10	0.0035	> 0.10
Passed?	No	Yes	Yes	Yes	No	Yes	No	Yes
P s.	**	ns	ns	ns	**	ns	**	ns
D-P t.								
P value	< 0.0001	0.5677	0.2930	0.4029	< 0.0001	0.5096	0.0002	0.5811
Passed?	No	Yes	Yes	Yes	No	Yes	No	Yes
P s.	***	ns	ns	ns	***	ns	***	ns
S-W t.								
P value	< 0.0001	0.6633	0.0044	0.1563	< 0.0001	0.3896	0.0006	0.4928
Passed?	No	Yes	No	Yes	No	Yes	No	Yes
P s.	***	ns	**	ns	***	ns	***	ns

4.5 Frequency of markers

This section will focus on the frequency of each marker. Tables 85 and 86 show a list of every marker used in the corpora along with their frequencies normalised per million. The reason we are changing now to a million normalisation is because we are comparing total numbers in the corpora instead of counts per speaker.

Table 85: Frequency of each marker in the Spanish corpus

AUX		**ADV**		**ADJ**		**MOOD**	
Marker	**Freq.**	**Marker**	**Freq.**	**Marker**	**Freq.**	**Marker**	**Freq.**
Poder	3813.1	A lo mejor	524.3	Seguro	119.5	Imperative	1380.6
Ir a	3152.7	Quizás	152.7	Posible	73.0	Subjunctive	106.2
Tener que	2246.7	Probablemente	102.9	Imposible	49.8		
Haber que	736.7	Seguramente	73.0	Obligado	26.5		
Deber	341.8	Obviamente	43.1	Necesario	13.3		
Haber de	6.6	Indudablemente	33.2	Prohibido	6.6		
		Posiblemente	29.9	Obvio	6.6		
		Sin duda	23.2	Preferible	3.3		
		Necesariamente	19.9	Obligatorio	3.3		
		Tal vez	13.3				
		Sin falta	6.6				
		Indefectiblemente	3.3				

4.5.1 Auxiliary usage

The most frequent auxiliary in both languages is semantically the same, the construction *Poder* + V and できる (*dekiru*). They can be translated to English as 'can', 'could', 'may' or 'might'. The Spanish marker carries the ambiguous meaning of epistemic modality (possibility of the state of affairs) and deontic (capacity or

Table 86: Frequency of each marker in the Japanese corpus

AUX		ADV		ADJ		MOOD	
Marker	**Freq.**	**Marker**	**Freq.**	**Marker**	**Freq.**	**Mark.**	**Freq.**
できる *dekiru*	1315.8	多分 *tabun*	1339.3	確か *tashika*	188.0	Potential	454.3
たい *tai*	1065.2	絶対 *zettai*	329.0	必要 *hitsuyō*	70.5	Imperative	70.5
なければならない *nakereba-naranai*	595.3	きっと *kitto*	195.8	無理 *muri*	70.5		
かも *kamo*	532.6	あるいは *aruiha*	141.0	可能 *kanō*	7.8		
ください *kudasai*	485.6	必ず *kanarazu*	141.0				
ればいい/たらいい *rebaii/taraii*	368.1	是非 *zehi*	125.3				
だろう *darō*	313.3	恐らく *osoraku*	47.0				
てもいい *temoii*	219.3	ひょっとしたら *hyot-toshitara*	7.8				
方がいい *hōgaii*	117.5						
ほしい *hoshii*	86.2						
ざるを得ない *zaruwoenai*	23.5						
しかない *shikanai*	23.5						
もらいたい *moraitai*	23,5						
はず *hazu*	23,5						
すべき *subeki*	15,7						
つもり *tsumori*	15.7						
わけにはいかない *wakenihaikanai*	7,8						

ability, and permission), whereas the Japanese can only be used in a deontic situation (capacity or ability). In terms of negation, both can change from a possibility to a necessity if negated as seen in the following examples 71[3] and 72[4] taken from the corpora:

(71) a. *... si Loren y Yoli no pued-en ven-ir lo que*
... if Loren and Yoli NEG can-PRES.MODAUX come-INF what CONJ
hace-mos
do-PRES

'... If Loren and Yoli can't come what we do...'

b. si Loren y Yoli <w neg="yes">no</w> <m class="AUX" modtype="NEC" neg="yes" subtype="AMBG" value="0%">pueden venir</m> lo que hacemos</Utterance>

(72) a. 僕 も うまく 説明-でき-ない ん だけど、
boku mo umaku setsumei-deki-nai n dakedo
definitely INCL well explain-can-NEG.MODAUXNEG EXPL however

'I can't explain it very well either...'

b. 僕もうまく <m class="AUX" modtype="NEC" neg="yes" subtype="DEON" value="0%"> 説明できない </m> んだけど、

In Spanish, *poder* is not only the most frequent auxiliary but also the only one that can mark a possibility in its non-negated form, which may explain its high frequency. The rest of auxiliaries would only mark necessity when they are not negated. The second most used auxiliary closely following *poder* is the periphrasis

[3]UNIT id: 5812 of the corpus. Speaker: YOV
[4]UNIT id: 8275 of the corpus. Speaker: CHO

Ir a ('will'), which implies an intention of the subject of the message. In terms of frequency it is succeeded by *tener que* ('have to', 'must') in third place. In Japanese, the second most frequent auxiliary is the construction たい (*tai*), a desire or, sometimes, implied intention of the subject, followed by necessity なければならない (*nakereba naranai*) ('have to', 'must') and general possibility (かも, *kamo*, 'may', 'might') in fourth place.

These three pairs of most frequent auxiliaries in both Spanish and Japanese seem to be roughly semantically equivalent, suggesting that the most used notions by speakers in both spoken languages are related to ability, capacity, desire, intention and obligation. As we have seen earlier, Spanish *poder*, which can express a general possibility but also a capacity or an ability, is equivalent to Japanese auxiliaries できる (*dekiru*) and かもしれない (*kamoshirenai*), the first and fourth most frequent markers, respectively.

Auxiliary *ir a* in Spanish involves an intention of the subject as example 73[5] shows, but may also imply a desire, more clearly seen in example 74[6], especially if used by the speaker in a first person, which is closer to Japanese *tai* (example 75). The closest marked modal auxiliary to Spanish *ir a* in Japanese that involves an intention of the subject is つもり (*tsumori*). However, the usage of this marker in the corpus is extremely low, which suggests that intention is Japanese would mostly be encoded using the non-marked, present form.

(73) a. *te v-oy a regal-ar un montón de cosas porque*
I will-PRES.MODAUX to.CONN give-INF a lot of things because
teng-o un montón de
have-PRES a lot of

'I'll give you a lot of things because I have a lot of...'

b. te <m class="AUX" modtype="NEC" neg="no" subtype= "DEON" value="100%">voy a regalar</m> un montón de cosas porque tengo un

[5]UNIT id: 2983 of the corpus
[6]UNIT id: 4642 of the corpus

montón de

(74) a. *porque no y no me v-oy a*
because no and not.AUXNEG I will-PRES.MODAUX to.CONN
dedic-ar a program-ar porque no me gusta
work-INF at program-INF because not I like

'but no, I won't work (i.e. I don't want to work) at programming because I don't like it.'

b. porque no y <w neg="yes">no</w> me <m class="AUX" modtype="NEC" neg="no" subtype="DEON" value="100%">voy a dedicar</m> a programar porque no me gusta

(75) a. それで おばあちゃん は その 時に 行-き-た-くない
sorede obāchan wa sono tokini i-ki-ta-kunai
so grandma NOM that occasion go-CONT-want-NEG.MODAUX
ん だけど、
n dakedo
EXPL but

'So (my) grandma does not want to go on that occasion but,'

b. それでおばあちゃんはその時に <m class="AUX" modtype="NEC" neg="yes" subtype="DEON" value="0%">行きたくない</m> んだけど、

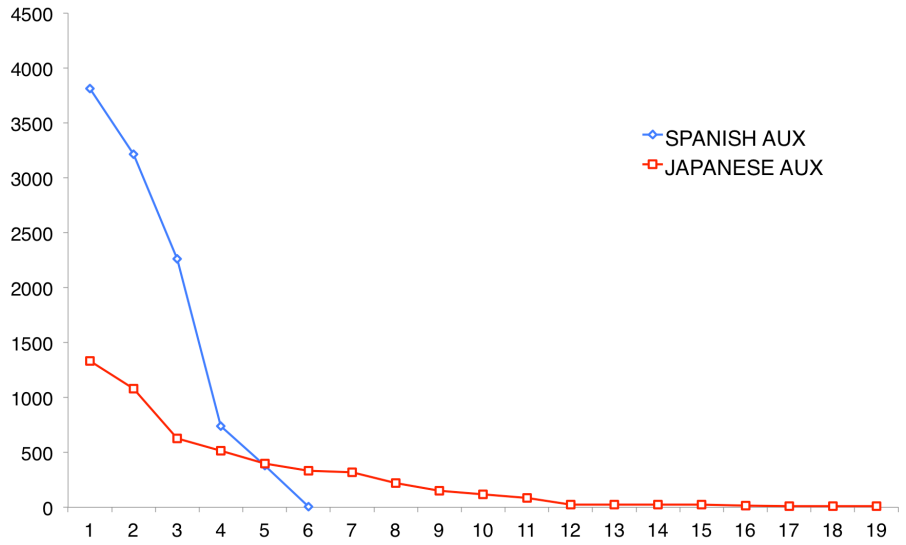
Comparing these results to other quantitative studies, Gómez Manzano's research (1991, p. 213) research reveals that *poder* + V is also the most frequent periphrasis in her spoken corpus, followed by *ir a*, *tener que*, *haber que*, *deber* and *haber de*, the same results as the ones achieved in this study (see Table 85). The

frequencies in this study, however, are not normalised, and we cannot perform a proper comparison with the ones from the C-ORAL-ROM corpus. In Japanese, Narrog (2009a, p. 168)'s study reveals that *tai* is the most frequent auxiliary, followed by *nakerebanaranai* and very closely by *dekiru* in third place and *kamoshirenai* in fourth (among the same selected markers). Just as before, the fact that these frequencies are not normalised, and they include both spoken and written data, does not allow us to perform proper comparisons. Nevertheless, the idea that the notions of ability, general possibility, desire, intention and obligation are the most frequent ones in both languages seems to be sustained.

Observing the least frequent auxiliaries in Japanese, we can find among the lower range the ones that suggest a necessary state of affairs, that the message is unavoidable and must be realised, such as ざるを得ない (*zaruwoenai*), しかない (*shikanai*) or はず (*hazu*). Japanese speakers appear to avoid these markers, selecting なければならない (*nakerebanaranai*) as the preferred general necessity marker, or using adverbs for this purpose, as we will see below. Comparing the equivalent Spanish markers that imply an unavoidable SOA, *Tener que* is the preferred one, with a considerably higher usage than *Haber que* or *Deber*, although this marker can also be used for epistemic purposes.

Figure 46 represents the progression of the frequencies of the auxiliaries in both corpora. The slope of each curve shows the difference in both languages: Spanish has less auxiliaries and the difference in usage between them is very disperse, especially from the third marker. Japanese allows a wider array of markers, which have a smoother decline in their frequencies.

Figure 46: Comparison of the progression of auxiliary frequencies in spoken Spanish and Japanese (per mill.)



4.5.2 Adverb usage

Observing previous frequency tables (85 and 86), Japanese speakers make a higher use (nearly 3 times higher) of adverbs than Spanish speakers. The most frequent adverb in both languages with a considerable distance from the second one is the one that encodes an epistemic possibility ('maybe', 'perhaps'): *A lo mejor* in Spanish and 多分 (*tabun*) in Japanese (see examples 76[7] and 77[8]). The next in terms of frequency in Spanish are also possibility adverbs, *quizás* and *probablemente*, leaving the necessity adverbs in a lower position. In Japanese, however, the following adverbs are of the necessity type, like 絶対 (*zettai*), きっと (*kitto*) and かならず (*kanarazu*). Hence, contrary to what we saw in the auxiliaries, although the most frequent one is a possibility adverb, there is a wider range of necessity markers with higher frequencies in Japanese adverbs. Overt necessity appears to be used more freely with adverbs in this language.

[7]UNIT id: 4960 of the corpus
[8]UNIT id: 4836 of the corpus

(76) a. *que no sé si riz-ar=me el pelo porque teng-o las*
CONJ don't know if curl-INF=CLTC the hair because have-PRES the
puntas fatal tía pero fatal y a lo mejor sí me lo
ends terrible mate just terrible and maybe.MODADV yes I it
riz-o
curl-PRES

'Well I don't know if I should curl my hair mate because the ends are terrible just terrible and maybe I do curl it'

b. que no sé si rizarme el pelo porque tengo las puntas fatal tía pero fatal y <m class="Adverb" modtype="POSS" neg="no" subtype="EPIS" value="50%" >a lo mejor</m> si me lo rizo

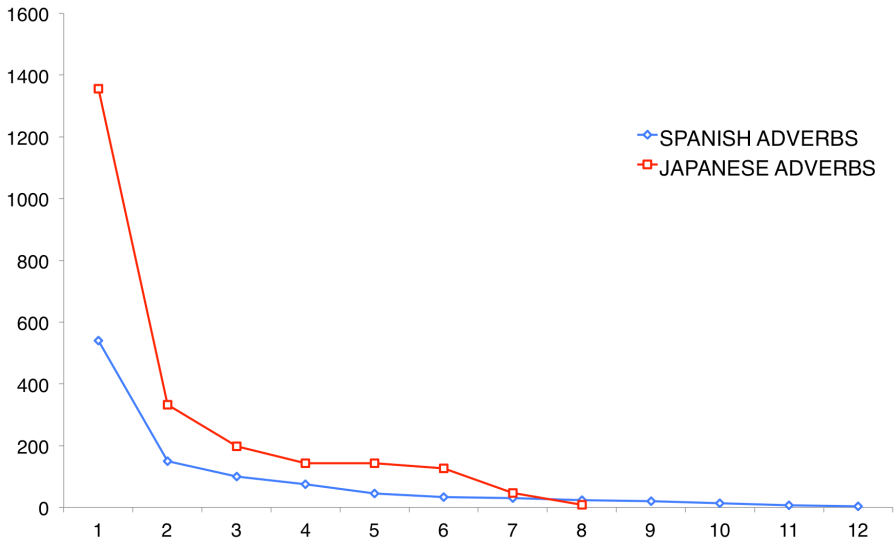
(77) a. 多分 その 日本人 も その 人 が 外国人
tabun sono nihonjin mo sono hito ga gaikokujin
maybe.MODADV that Japanese INCL that person NOM foreigner
だっていう 事 を 忘れ-て
datteiu koto wo wasure-te
regarding matter ACC forget-TE

'Maybe that Japanese(s) also forgets that person(people) is(are) a foreigner(s)'

b. <m class="Adverb" modtype="POSS" neg="no" subtype="EPIS" value="70%"> 多分 </m> その日本人もその人が外国人だっていう事を忘れて。

All in all, the relative usage of the modal adverbs in both languages appear to be rather similar, as seen in Figure 47: they begin with a very frequent adverb, more than twice of the frequency in the Japanese one, and then drop to a series of adverbs that very gradually reduce their quantities in an almost parallel way in both languages, with a slight advantage from the Japanese adverbs.

Figure 47: Comparison of the progression of adverb frequencies in spoken Spanish and Japanese (per mill.)



4.5.3 Adjective usage

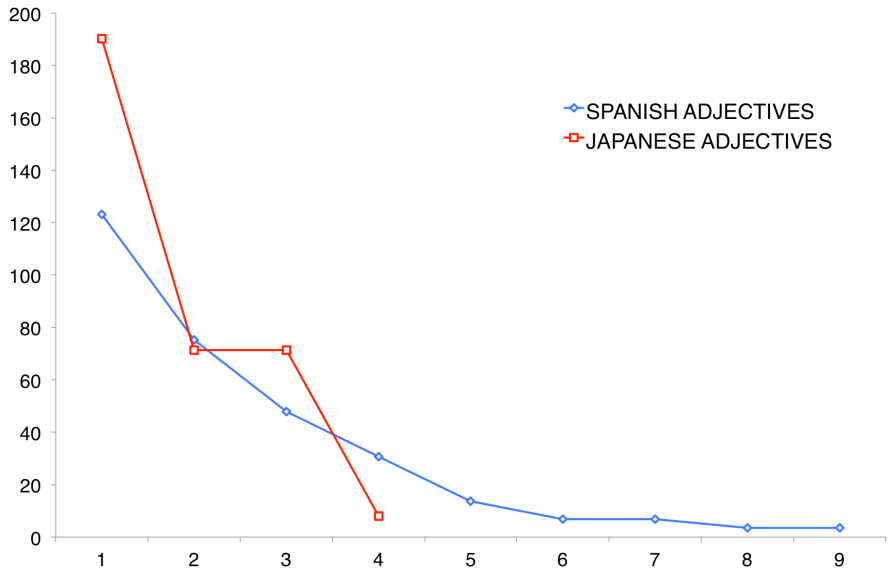
As with the adverbs, there appears to be a coincidence in the usage of modal markers between both languages: in Spanish and Japanese the most frequent predicative adjective denotes a necessity ('certain', 'sure'), *seguro* in the former, 確か (*tashika*) in the latter. However, there is a broader array of options used in Spanish. In Japanese, only four different modal adjectives can be found in the corpus, 確か, 必要 (*hitsuyō*, 'necessary'), 無理 (*muri* 'impossible') and 可能 (*kanō* 'possible'). The last one, the only one that denotes a possibility, is scarcely used. As with the adverbs, Japanese adjectives are primarily for implying necessity, with more different options.

In Spanish, the majority of adjectives also encode necessity modality (with 8 different possibilities). There is only one that encodes possibility, but it is in second place in terms of frequency. Therefore possibility can be frequently marked in adjectives, contrary to Japanese, but it seems to be limited to only one adjective as well.

There is clearly a very marked difference among the relative usage of modal adjectives (Figure 48). The higher number of options in Spanish create a frequency curve without very substantial differences in the frequencies. However, the four

adjectives used in the Japanese corpora offer a very different picture: the first one has a very high frequency, followed by two with the same number, and finished with a very low one in the end. The picture is very similar to the one seen in the auxiliaries (Figure 46), although the languages have been reversed.

Figure 48: Comparison of the progression of adjective frequencies in spoken Spanish and Japanese (per mill.)



4.5.4 Mood usage

Probably the highest difference between Spanish and Japanese can be found in the usage of imperative mood. Spanish imperatives, and their negative counterpart negating the subjunctive form of the verb, add up to nearly 1400 instances per million words. In Japanese, the imperative is only used in a total of 70 words per million. Moreover, nearly all the usages of the imperative in the C-ORAL-JAPÓN corpus are used in a quoted manner by the speaker, referencing to an imperative told by a third person. Therefore, the necessity marker is not used directly upon the hearer.

Spanish excels in its usage of overt necessary imposition of the SOA on the receiver of the message through mood, whereas Japanese speakers avoid it, probably due to its excessive directness. Authors in Kaiser et al. (2013, p. 492) consider that this form is limited to certain situations such as traffic signs, robbers, slogans, etc.

and male speakers in informal communications. The higher frequency in Spanish shows that it is a form that can be used much more freely.

If Japanese want to transmit a necessary SOA to the hearer, other modal markers are used instead, such as necessity auxiliaries. Looking back at Table 86, there is a wide number of Japanese modal auxiliaries with fairly high frequencies that may be used instead of the imperative for impositions or requests, such as ください (*kudasai*), ればいい/たらいい (*rebaii/taraii*), た方がいい ((*tahōgaii*) or すべき (*subeki*) (see Examples 78[9], 79[10], 80[11]). Spanish auxiliaries *tener que*, *deber*, the negative usage of *poder* and even *haber que* can also be used for this matter, especially if combined with the conditional form of the auxiliary and the second person (see Examples 81[12], 82[13], 83[14]), although their usage in the corpus is not so high. Nevertheless, the high amount of imperatives and negative subjunctives in its corpora suggests that their usage has not been replaced by auxiliaries as can be the case in Japanese.

(78) a. それで それだけ 歩-ければ あと は ほとんど 連れ-て-

sorede *soredake* *aru-kereba* *ato* *wa* *hotondo* *tsure-te-*

so that much walk-COND after NOM mostly accompany-TE-

-ください-ます

-kudasai-masu

-please-POL.MODAUX

'If you can walk that much, please accompany me for the rest'

b. それでそれだけ歩ければあとはほとんど <m class="AUX" mod-type="NEC" neg="no" subtype="DEON" value="100%"> 連れてくださいます </m> から。

[9]UNIT id: 13948 of the corpus
[10]UNIT id: 17224 of the corpus
[11]UNIT id: 13372 of the corpus
[12]UNIT id: 14788 of the corpus
[13]UNIT id: 14286 of the corpus
[14]UNIT id: 12023 of the corpus

(79) a. うん 食堂 ある から そこで 食べ-ればいい 工場
un *shokudō* *aru* *kara* *sorede* *tabe-reba-ii* *kōjō*
yes cafeteria is since there eat-COND-good.MODAUX factory
の 近く に あっ-た から
no *chikaku* *ni* *at-ta* *kara*
GEN near LOC is-PAST since

'Yes, since there is a cafeteria you should eat there, it is near the factory'

b. うん食堂あるからそこで <m class="AUX" modtype="NEC" neg="no" subtype="DEON" value="100%"> 食べればいい </m>、工場の近くにあったから。

(80) a. 文化祭 は やっぱり あっ-た-ほうがいい よ
bunkasai *wa* *yappari* *at-ta-hōgaii* *yo*
cultural festival NOM also be-PST-should.MODAUX EMPH

'There should also be a cultural festival'

b. 文化祭はやっぱり <m class="AUX" modtype="NEC" neg="no" subtype="DEON" value="100%"> あったほうがいい </m> よ。

(81) a. *sí tien-es que practic-ar mucho para llegar a ser*
yes have-PRES.MODAUX to.CONN practise-INF a lot in order to be
como Michael Schumacher
like Michael Schumacher

'Yes you should practise a lot in order to be like Michael Schumacher'

b. sí <m class="AUX" modtype="NEC" neg="no" subtype="AMBG" value="100%">tienes que practicar</m> mucho para llegar a ser como Michael Schumacher

(82) a. *porque incluso si quier-es lo que pasa que claro*
because even if want-PRES what CONJ happens CONJ of course
ten-drías que mir-ár=te=lo muy bien
have-COND.MODAUX to.CONN look at it-INF=CLTC=CLTC very closely
por si te pregunt-an pero est-ás
in case if cltc ask-PRES but are-PRES

'Because even if you want what happens is... of course, you should look at it very closely in case they ask you but you are'

b. porque incluso si quieres lo que pasa que claro <m class="AUX" modtype="NEC" neg="no" subtype="AMBG" value="100%">tendrías que mirártelo</m> muy bien por si te preguntan pero vamos estás

(83) a. *y eso ha-bría que repercutir=lo*
because even have-COND.MODAUX to.CONN cause an effect=CLTC.INF
directamente a los clientes claro
want what CONJ happens CONJ

'And we should cause an effect of that over the clients, of course'

b. y eso <m class="AUX" modtype="NEC" neg="no" subtype="DEON" value="100%">habría que repercutirlo</m> directamente a los clientes claro

4.6 Modification of modal markers in the spoken discourse

This part of the study will tackle the elements and features that may affect the modal markers in the spoken discourse: negation, ellipsis, separation and misspellings of modal markers in both languages.

4.6.1 Negation

Observing the frequencies of the Spanish negative elements that appear with modal markers, represented in Table 87, all of the most typical negative adverbs are used in the corpora (*no*, *tampoco*, *nunca*) with the exception of *jamás*. The most frequent negative particle is adverb *No* ('no'), with a frequency of 585 apparitions, nearly 40 times more frequent than the next marker, *Tampoco* ('neither', 'nor'). This is not surprising, as it is considered the most characteristic negative adverb in Spanish (RAE, 2009, p. 3632), allowing it to appear with any type of auxiliary or adjective.

Table 87: Negative elements modifying Spanish modal markers

Negative Element	**Raw Frequency**	**Normalised (per mill.)**
No	585	1941.40
Tampoco	15	49.78
Ni	9	29.87
Nunca	7	23.23
Nadie	2	9.96
Difícilmente	2	6.64
Nada	1	3.32
Sin	1	3.32

Tampoco, formed by the union of *tan* ('so') and *poco* ('few'), obtains the second place in terms of quantity of usage, followed by the conjunction *ni* ('even') and ad-

verb *nunca* ('never') (Example 84[15]). In the corpus, they appear with the auxiliaries *poder* and *ir a*. That is, they can either modify the auxiliary verb, changing the type of modality, or affect only the main verb as in the latter, maintaining the necessity.

(84) a. *al día siguiente ni la pod-ía mir-ar me entr-aba*
on the next day NEG her can-PST.MODAUX look-INF cltc feel.PST
unas arcadas Pili
a gag reflex Pili

'I couldn't look at her next day, Pili, it made me sick'

b. al día siguiente <w neg="yes">ni</w> la <m class="AUX" modtype="NEC" neg="yes" subtype="AMBG" value="0%">podía mirar</m> me entraba unas arcadas Pili

The distance of the negative elements from the marker they are modifying has been summarised in Table 88. The maximum distance recorded in the corpus is of two words, and the minimum is zero. An example of distance one can be found in example 85[16]. The negative 'no' (*not*) is separated from the periphrastic construction 'puede cambiar' (*can change*) by the pronoun 'se' (*it/him/herself*), and therefore the distance is 1.

(85) a. *pero es que eso no se pued-e cambiar*
but it's just that that NEG it can-PRES.MODAUX look.INF

'But, it's just that, it cannot be changed'

b. pero es que eso <w neg="yes">no</w> se <m class="AUX" modtype="NEC" neg="yes" subtype="AMBG" value="0%">puede cambiar</m>

[15]UNIT id: 4347 of the corpus
[16]UNIT id: 5335 of the corpus

Table 88: Word distance between negative elements and Spanish modal markers

Negated markers	**Max. distance**	**Min. distance**	**Average dist.**
436(11,24% of total)	2	0	0,42

Looking now at the Japanese negative elements in Table 89, the plain negative auxiliary ない (*nai*) is, as Spanish *no*, the most frequent one with 87 instances, with a considerable distance from the polite form ません (*masen*), with only 8 usages. Its variations appear to be less frequent. There are 6 cases of the shortened version ん (*n*), 3 instances of はない *hanai* (from which only 1 has been used as the longer ではない (*dehanai*) version) and only 1 of じゃない (*janai*). We can also find one sentence (see Example 86[17]) of the construction なくなる (*nakunaru*), lit. 'become not'. This has been used along the auxiliary できる (*dekiru*, 'can', 'may') to create a negative possibility: 'it cannot become something':

(86) a. 我慢-でき-な-くなっ-た　　とき　が　あっ-て、
gaman-deki-na-kunat-ta　　*toki*　*ga*　*at-te*
endurance-can-become-NEG-PST.MODAUXNEG　when　NOM　ISTE

'When I couldn't endure (it),'

b. <m class="AUX" modtype="NEC" neg="yes" subtype= "AMBG" value="0%"> 我慢できなくなった </m> ときがあって、

[17]UNIT id: 8149 of the corpus

Table 89: Negative elements modifying Japanese modal markers

Negative Element	Raw Frequency	Normalised (per mill.)
ない (*nai*)	87	681.41
ません (*masen*)	8	62.66
ん (*n*)	6	46.99
はない (*hanai*)	3	23.50
じゃない (*janai*)	1	7.83
なくなる (*nakunaru*)	1	7.83

4.6.1.1 Negated markers

The following Table 90 shows which modal markers have been negated in the Spanish corpus, their frequency normalised per million words and the percentage from the total. In other words, for example, the number of negated *Ir a* auxiliaries is the 11,60% of the total of apparitions of *Ir a*, the number of negated necessity adjectives is the 0,31% of the total of necessity markers, and the total number of negated subjunctive is 100%, as it is the way the negative imperative is formed.

Overall, the most frequent negated Spanish modal marker are auxiliaries, although hardly surprising at it is the most frequent marker. Nevertheless, the proportions are fairly high: more than 13% of the necessity modality in the corpus is formed by negated auxiliaries, nearly 13% of all the auxiliaries are negated, and more than 20% of all the *Poder* periphrases, the only ones that encode possibility and change when negated, undergo some kind of negation. They are followed by *Ir a* and *Deber*, with around 12% of negated forms, although in these cases necessity is maintained, moving to a 0% of probability value instead. These numbers contrast with the low results in possibility modality: only a 3% of it is negated through auxiliaries.

Negation is also high between predicative adjectives. More than 20% of them are negated, which represent the 0,31% and the 0,95% of all the possibility and

necessity, respectively. The most frequent is *posible* ('Possible'), which receives both the morphological negation (47,88 cases per mill.) and the lexical (10,26).

Table 90: Negated Spanish modal markers

Class	**Negated Element**	**Freq(per mill.)**	**% from total**
AUX	Poder	776.56	20.36 (Poder)
	Ir a	368.37	11.60 (Ir a)
	Tener que	86.28	3.93 (Tener que)
	Deber(de)	39.82	10.81 (Deber)
	Haber que	33.18	4.63 (Haber que)
	TOTAL	1304.22	12.61 (Auxiliaries)
	NEC	1191.38	13.36 (Necessity)
	POSS	112.83	2.68 (Possibility)
Subjunctive	eches	9.95	
	pases	9.95	
	fíes	6.64	
	hables	6.64	
	vayas	6.64	
	pongas	6.64	
	solapes	6.64	
	bañes	3.32	
	comas	3.32	
	creas	3.32	
	cuentes	3.32	
	desprecies	3.32	
	desvíes	3.32	
	digas	3.32	
	jorobes	3.32	
	llames	3.32	
	marees	3.32	
	pintes	3.32	
	quedes	3.32	
	quejes	3.32	
	quieres	3.32	
	seáis	3.32	
	vengas	3.32	
	TOTAL	106.20	100.00 (Subjunctive)
	NEC	106.20	1.22 (Necessity)
Adjective	Posible	59.72	
	Obligado	6.63	
	Necesario	3.31	
	TOTAL	69.69	22.22 (Adjectives)
	NEC	26.55	0.31 (Necessity)
	POSS	43.14	0.95 (Possibility)

Table 91 shows the same case with Japanese modal markers that have been negated. As in Spanish, the most frequent are the negated auxiliaries, which also comprise almost the 13% of all the necessity modality in the corpus. The Japanese equivalent to *Poder*, できる (*dekiru*), is also the highest negated auxiliary, with a 42%. That is, more than 40% of all the cases of *dekiru* are in negated form. They are followed by the desiderative and petition forms たい (*tai*) and てくれる (*tekureru*).

The second most frequent negated type of marker is the potential mood which, as we described before, is very near to be the equivalent to *dekiru*. 23% of this mood form is being negated in the corpus, although it only represents a 2% of the total necessity modality used.

Finally, the only predicative adjective that has been found in a negative state in the corpus is 必要 (*hitsuyō*, 'necessary'), although the amount is fairly high as it represents the 9% off all the predicative adjectives.

Table 91: Negated Japanese modal markers

Class	**Negated Element**	**Freq(per mill.)**	**% from total**
AUX	できる (*dekiru*)	595.25	42.26 (*dekiru*)
	たい (*tai*)	101.82	9.56 (*tai*)
	TOTAL	697.07	13.32 (Auxiliaries)
	NEC	697.07	14.73 (Necessity)
Potential	覚えられる (*oboerareru*)	15.66	
	かける (*kakeru*)	7.83	
	とがめる (*togameru*)	7.83	
	よけられない	7.83	

	借りられる (*karirareru*)	7.83	
	出られる (*derareru*)	7.83	
	受け入れる (*ukeireru*)	7.83	
	来られる (*korareru*)	7.83	
	決められる (*kimerareru*)	7.83	
	着けられる (*tsukerareru*)	7.83	
	答えられる (*kotaerareru*)	7.83	
	考えられる (*kangaerareru*)	7.83	
	触れられる (*furerareru*)	7.83	
	食べられる (*taberareru*)	7.83	
	TOTAL	109.65	23.73 (Potential)
	NEC	109.65	2.34 (Necessity)
Adjective	必要 (*hitsuyō*)	31.32	
	TOTAL	31.32	9.30 (Adjectives)
	POSS	31.32	0.84 (Possibility)

4.6.2 Ellipsis of modality

In the spoken discourse, especially the spontaneous and informal one, speakers seldom tend to minimise the amount of uttered information, leaving out unnecessary or redundant elements (Briz Gómez, 2001, p. 83). Also, spoken language is much more context-bound than written language. The participants in a spoken communicative interaction share a common knowledge (either mentioned before in the exchange, or through general understanding of the world) (McCarthy, 1998, p. 64) which makes them drop the unnecessary words. The dropped words are considered to be able to recover given information, such as in the sentences 'I was ill yesterday. So was my wife' (Miller & Weinert, 1998, p. 209). We will consider the following criteria for an elliptic element (Quirk et al., 1985):

1. The ellipted word is recoverable
2. The elliptic construction is ungrammatical
3. The insertion of missing words results in a grammatical sentence
4. The dropped elements are textually recoverable
5. The dropped elements are present in the text

Nevertheless, some of these conditions may not be true in some cases of spontaneous, spoken discourse. For example, one of the features we can encounter are fragmented sentences, utterances that have been cut or left unfinished by an interruption of the speaker's utterance. This may leave the dropped element unrecoverable. An example of this can be seen in the following sentence[18] (87). The speaker starts a modal periphrastic construction but stops after the auxiliary, without adding the main verb anywhere in the following text:

(87) a. *la cocina también cuando te pued-es*
the kitchen too when you can-PRES.MODAUXELLI

'The kitchen too when you can'

[18]UNIT id: 2573 of the corpus

b. la cocina también cuando te <m class="AUX" elli="yes" modtype= "POSS" neg="no" subtype="AMBG" value="50%">puedes<v_ elli type= "inf"/></m>

In some areas of applied linguistics such as language teaching and acquisition, some studies have revealed that regular elliptic operations in conversations such as subject-dropping are almost limited to native speakers –Scarcela & Brunak (1981), seen in McCarthy (1998). For automatic data retrieval, it can also be a problem, especially when processing spoken texts, as the ellipted element could be found in a previous utterance, or even worse, not mentioned in the portion of the transcription. For this study, the main issue comes with the dropping of one of the elements in a multiword modal construction such as a periphrasis. The main verb can be dropped leaving only the auxiliary, or vice-versa, as seen in previous example 87 and the following 88[19]:

(88) a. できｰない と 思ｰう、
deki-nai *to* *om-ō*
can-NEG.MODAUXNEGELLI QUOT think-PLN

'I think (you) cannot'

b. <m class="AUX" elli="yes" modtype="NEC" neg="yes" subtype= "AMBG" value="0%"><v_elli type="inf"/> できない </m> と思う、

The following tables, Tables 92 and 93, show the auxiliaries that contain an elliptic element in the corpora. Once again, Spanish *Poder* and Japanese できる ('can', 'may') occupy the first place in terms of frequency usage, although the Japanese auxiliary is the only one that have undergone an elliptic process.

In terms of proportion with the rest of overall modality markers, in both languages the elliptic process is fairly rare. In Spanish, elliptic auxiliaries form only the

[19]UNIT id: 11112

1% of the total modal markers, whereas in Japanese this number goes up to nearly 4%, although they only occur with marker できる (*dekiru*, 'can').

Table 92: Spanish auxiliaries with elliptic elements

Class	**Negated Element**	**Freq(per mill.)**	**% from total**
AUX	Poder	92.92	2.33 (Poder)
	Haber que	9.95	1.39 (Haber que)
	Tener que	9.95	0.45 (Tener que)
	Ir a	6.63	0.11 (Ir a)
	TOTAL	119.46	1.08 (Auxiliaries)
	NEC	89.60	0.96 (Necessity)
	POSS	29.86	0.63 (Possibility)

Table 93: Japanese auxiliaries with elliptic elements

Class	**Negated Element**	**Freq(per mill.)**	**% from total**
AUX	できる (*dekiru*)	206,12	15,4 (*Dekiru*)
	TOTAL	206,12	3,89 (Auxiliaries)
	NEC	63,42	1,34 (Necessity)
	POSS	142,70	3,77 (Possibility)

4.6.3 Separation of modality

Another feature that should be taken into account when creating the automatic modality tagger is the possible separation in the discourse of the main and auxiliary verb in periphrastic constructions. It is treated similarly to the ellipsis, but in this case both elements are present in the same utterance (See Example 89[20]). Also, as with negation, the average distance between auxiliary and main verbs has been recorded, summarised in Table 94:

[20]UNIT id: 11112

(89) a. *...es decir que el abogado le pod-ría pues*
that is that the lawyer him can-COND.MODAUXID_1 well
inform-ar sobre otras personas...
let know-INFREF_1 about other people

'[...] that is, the lawyer can, well, let him know about other people [...]'

b. [...] es decir que el abogado le <m class="AUX" id="1" modtype="POSS" neg="no" subtype="AMBG" value="50%">podría</m> pues <m class="AUX" modtype="POSS" neg="no" ref="1" subtype="AMBG">informar </m> sobre otras personas [...]

Table 94: Word distance between auxiliary and main verbs in Spanish and Japanese

Language	**Separated markers**	**Max. dist.**	**Min. dist.**
Spanish	194,93(1,46% of total)	4	1
Japanese	15.66(0,18% of total)	2	1

The maximum distance possible between the main verb and its auxiliary in Spanish is 4 words, compared to the maximum of 2 words in Japanese (See Example 90[21]). Nevertheless, the proportion compared to the total modal markers is minimal, not reaching a 0,5% (only three cases have been found in the Japanese corpus, approximately 39 per million words).

[21]UNIT id: 13614

(90) a. 立つ-ということ は あのう でき-ない
tatsu-toiukoto *wa* *anō* *deki-nai*
stand up.PLN-NMZ.REF_1 NOM well can-NEG.MODAUXNEGID_1
ん です よ ね
n *desu*COP *yo* *ne*
EXPL EMPH INT

'(you) cannot quite, well, such thing as standing up, right?'

b. <m class="AUX" id="1" modtype="NEC" neg="yes" subtype="DEON" value="0%"> 立つということは </m> なかなかあのう <m class="AUX" ref="1" modtype="NEC" neg="yes" subtype="DEON" value="0%"> できないん </m> ですよね

The majority of words that could be introduced inside a Spanish periphrasis are either repetitions of the conjunctions 'a' or 'que' by stammering made by the speaker (e.g. *... tengo que que hacer...* '... I have to to do...'), personal pronouns, adverbs or exclamations/discourse markers generally used to create a pause, hesitation in the utterance (O'Connell & Kowal, 2005), as seen previously in Example 89. In Japanese we also find hesitation discourse markers like in the previous example (あのう, *anō* 'well') or the emphatic adverb ちゃんと (*chanto* 'seriously, perfectly').

Lastly, Tables 95 and 96 show the elements separated form the auxiliary verb in Spanish and Japanese respectively, as well as their proportions from the total amount of auxiliaries, and necessity and possibility modality. It appears that auxiliaries that indicate necessity tend to undergo separation more easily, although the numbers are still too low to draw solid conclusions.

Table 95: Separated elements from the Spanish auxiliaries

Marker Class	Negated Element	Freq(per mill.)	% from total
AUX	a ver	10,26	
	decir	10,26	
	ir	6,84	
	que ir	6,84	
	a colaborar	3,42	
	a comer	3,42	
	a comprobarlo	3,42	
	a levantar	3,42	
	a sacar	3,42	
	a saludar	3,42	
	a trabajar	3,42	
	abordar	3,42	
	aceptar salir	3,42	
	adaptarla	3,42	
	amparar	3,42	
	aprobar	3,42	
	conseguir	3,42	
	concienciar		
	conservar	3,42	
	decirle	3,42	
	determinar	3,42	
	discutir	3,42	
	diseñar	3,42	
	encontrar	3,42	
	entrar	3,42	
	estar	3,42	
	haber comentado	3,42	
	hablando	3,42	
	llevar	3,42	
	luchar	3,42	
	modificar	3,42	
	objetar	3,42	
	pagar	3,42	
	pedir	3,42	
	poner	3,42	
	ponerte	3,42	
	preservar	3,42	
	presupuestar	3,42	
	que	3,42	
	que despedir	3,42	
	que entregar	3,42	
	que pedir	3,42	
	que poner	3,42	
	que renovarlo	3,42	
	quitar y poner	3,42	
	reducir	3,42	
	respetar	3,42	
	tener	3,42	
	TOTAL	181,26	1,74
	NEC	140,22	1,57
	POSS	41,04	0,95

Table 96: Separated elements from the Japanese auxiliaries

Marker Class	Negated Element	Freq(per mill.)	% from total
AUX	そういうの (*sōiuno*)	7,93	
	その人 (*sonohito*)	7,93	
	ウォーキングを (*uōkinguwo*)	7,93	
	攻めようというのが (*semeyōtoiunoga*)	7,93	
	立つということは (*tatsutoiukotoha*)	7,93	
	TOTAL	39,64	0,75
	NEC	39,64	0,84

4.6.4 Errors in modality

The last feature to take into consideration when observing and drawing possible rules and patterns for modal markers, especially in spoken discourse, is the possibility of speakers breaking these rules. Whether or not they should be considered *errors* with a negative connotation, or a positive instance of a possible language change, is outside the scope of this study. Spoken language is a very much alive entity, much more unstable and prone to changes than written discourse, most of the time triggered by *deviations* made in speech by native speakers (Tenfjord et al., 2006), which may become constant (Corder, 1967). Nevertheless, the objective of this study is to observe constants and generate rules for automatically tagging modal markers. Hence, an *error* will be considered as that which 'offends the norm (or standard) of the language' –Ringbom (1987), as seen in Gilquin & de Cock (2011), a feature that breaks the pattern observed in the majority with no negative connotation intended. For example, as we saw back in sentence 63, the usage of the verb in infinitive form instead of the imperative has become a fairly common feature in Spanish. Sentence 91[22] shows a triple repetition of this in a single utterance:

[22]ID: 1205

(91) a. *veniros* *y* *veniros* *llevaros*
come.MODIMP_ERR and come.MODIMP_ERR take.MODIMP_ERR
el *equipo*
the equipment

'Come and come, take away the equipment'

b. <m error="yes" class="mood_IMP" modtype="NEC" neg="no" subtype="DEON" value="100%">veniros</m> y <m error="yes" class="mood_IMP" modtype="NEC" neg="no" subtype="DEON" value="100%">veniros</m> <m error="yes" class="mood_IMP" modtype ="NEC" neg="no" subtype="DEON" value="100%">llevaros </m> el equipo

Table 97 shows which modal markers have been used erroneously by speakers in the corpus as well as their frequencies. It has only been studied in Spanish due to the lack of language level required for detecting possible errors in Japanese native speakers. The most frequent mistake takes place in the imperative mood, using the infinitive instead as explained previously. However, the frequency is quite low: only 3% of all the imperatives used have been replaced by the infinitive form. The next kind of error takes place in the auxiliaries, particularly between the necessity constructions *Deber* + V (deontic) and *Deber de* + V (epistemic), which many speakers use indistinctly, is much more serious: from all the instances of this construction, 36% have been wrongly used by the speakers.

Table 97: Markers that include errors made by native Spanish speakers

Marker Class	**Negated Element**	**Freq(per mill.)**	**% from total**
mood_IMP	llevar	6,84	
	sentar	6,84	
	venir	6,84	
	decir	6,84	
	aislar	3,42	
	estirar	3,42	
	imaginar	3,42	
	meter	3,42	
	mirar	3,42	
	TOTAL	44,46	3,19
	NEC	44,46	0,50
AUX	deber(de)	136,80	36,04
	ir a	3,42	0,11
	TOTAL	140,22	1,35
	NEC	140,22	1,57

4.7 Inferences from the quantitative study

These pages have presented a quantitative study of the usage of modal markers in two corpora, cross and intra-linguistically. We began with a series of hypotheses that have been evaluated and confirmed throughout the analysis:

1. In general, there is not a clear difference in the amount of modality used in Spoken Spanish and Japanese. Hypothesis rejected, the null hypothesis is confirmed: the difference between the means is significant and there is not a relation between them. Although, according to the type of interaction, the alternative hypothesis is accepted in monologues.
2. Necessity modality is significantly higher in Spanish. Hypothesis accepted, the difference is significant.
3. Deontic markers could be significantly less frequent than epistemic ones in Japanese. Again, the hypothesis is rejected in general numbers: deontics are higher, but loking at the type of interaction, it is only higher in monologues, were the threat of using FTAs is lower. In Spanish, the usage of deontic markers is extremely high compared to epistemic ones. Also, as expected, ambiguity in the subclassification achieves high numbers in Spanish, but is barely present in Japanese.
4. There a significant use of the auxiliary class. Hypothesis accepted: the tests show that modality marked with auxiliaries is significantly high in both languages in every situation.
5. Modality in general is significantly higher in interactive situations (dialogue/ conversations) than monologues. The hypothesis is accepted in both languages, although the difference is higher in Spanish.
6. Modality is dependent to the level of register, more frequent in informal situations. Hypothesis accepted, only it has only been confirmed in Spanish due to the lack of data in Japanese.
7. The amount of modal markers used depends on the gender of the speaker and there is variation among men and women. Hypothesis rejected: the difference is not significant in general. However, this only happens when looking at

possibility modality. Necessity modality, as well as deontic in Spanish and epistemic in Japanese, do show significant differences in gender.

8. The same was assumed with age. Hypothesis rejected: the tests showed a lack of noticeable variation according to age.
9. Finally, negation is highly frequent with no clear differences between both languages; ellipsis is lower than expected, but higher in Japanese; and separation of auxiliary and main verb is barely present but higher in Spanish.

These conclusions should be treated carefully as, afterall, this is but a limited example of language. However, the combination of descriptive and some inferential statistics in similar types of variables and data can provide us with an initial insight of the problem at hand, i.e. Spanish and Japanese modality (Baroni & Evert, 2008). With the automatisation, we can repeat rapidly the analysis in new, more extensive data in the future and check if these insights are sustained.

A first glance into the overall amount of markers in each corpus indicates that the usage is very different in both languages: the mean per speaker is significantly higher in Spanish than Japanese, and Japanese modality, which passes normality tests, is more contained and regular in terms of dispersion. This can throw down the possibility of finding similarities between both languages.

However, just as we keep dissecting the data, we do find common ground. Overall, the Spanish data is more irregular due to some spikes in the frequencies, but both languages appear to be similar in the lowest numbers. Following this, the necessity and possibility comparison draws the first noticeable similarity: where both languages differ most is in their usage of necessity markers, Spanish achieving higher and more dispersed data. Possibility markers, on the other hand, are very similar. The same situation is found in the grammatical class of the markers: both languages prefer significantly the auxiliaries.

A clearer picture is drawn when considering additional variables: modality is very similar and regular in monologues from both languages, which have almost identical proportions in the usage of necessity, possibility, epistemic and deontic markers. The differences strike out in dialogues: Spanish speakers remarkably prefer

necessity and deontic modality when interacting with others, whereas Japanese use nearly the same amount of necessity and possibility, and prefer epistemic markers when communicating.

Therefore, there are similarities in monologues and possibility modality, and differences in dialogues and necessity. When the communicative act is performed individually, both languages are quite similar, but in a confrontation, when the face of the speaker is threatened and extralinguistic elements come into play (such as social and cultural obligations and rituals), each language keeps its own path. Spanish speakers are less restrained, more different between each other, and overtly use necessity and deontic markers freely. Japanese, on the other hand, shut down when interacting with another participant, and limit the markers that could threaten the face of the other.

Also, women appear to use necessity more freely than men, and Spanish use more deontics and Japanese more epistemics than men. Nevertheless, the differences are not so noticeable, and joined with the lack of disparities in the age factor, reinforces the idea that modality is conditioned by the type of communication.

Another important feature extracted from the corpora is that modality's sub-classification into epistemic and deontic is not at all reliable. The ambiguous auxiliaries in Spanish, those that can carry either or both meanings, are among the most frequent. This jeopardises the comparative study and makes the epistemic/deontic frequencies not entirely reliable. But it also complicates the automatisation. A solution would be to manually disambiguate and apply machine learning, but the process could also be very challenging since the separation of both meanings is not so clear-cut, as we saw in Section 2.2.2. If we want to formalise Spanish modality and properly annotate it automatically for a quantitative study, it is best to remain in the necessity/possibility dichotomy, or look for a new classification.

The section dealing with the modification of markers is especially useful for the design of the tagger, but also complement these assumptions. The most frequent auxiliary, adverb and adjective is semantically equivalent in each language. The auxiliaries involve meanings of capacity, desire and intention, but Japanese use a

wider variety in the corpus, the vast majority used for deontic modality. Spanish, on the other hand, have higher variety in adverbs, encoding epistemic values. In other words, Japanese have at their disposal a higher array of options to express deontic actions, whereas in Spanish there are more ways of expressing epistemic notions. Also, negation is fairly frequent and constant: around the 13% of the auxiliaries in both corpora are modified by a negative element, and almost all of them imply a necessity. Ellipsis is not a common phenomena but the frequency is nearly double in Japanese than Spanish.

Chapter 5

Developing an automatic modality tagger

5.1 Description of the program

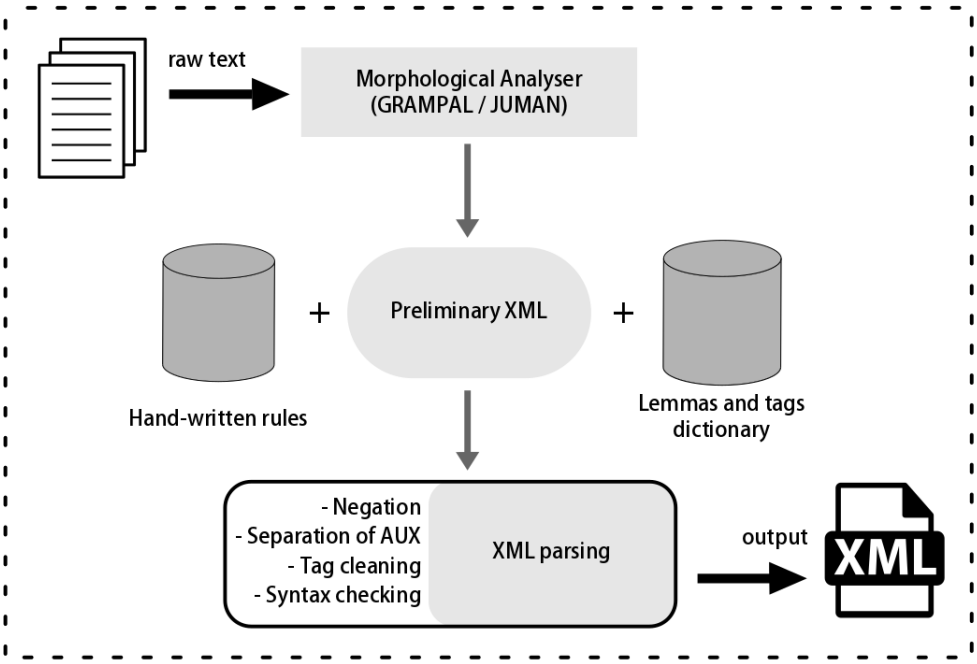
Up until now, we have discussed in this study the meaning and classification of modality and modal markers in Spanish and Japanese, their tagging in two corpora, and the quantitative results extracted from them. This section will describe how this knowledge has allowed the development of a Python program that would detect and classify modal markers in a Spanish or Japanese text.

The program has been designed with a series of ideas in mind:

1. It must be equally designed for Spanish and Japanese.
2. It must read a text, find the possible modal markers, and classify them according to the first (necessity/possibility) and second (epistemic/deontic/ambiguous) levels of modality, and assign it a class (auxiliary, adverb, adjective or mood).
3. It must tag the negation of modal markers and update their classification (if, for example, a possibility turns into a necessity).
4. It must overcome the possible challenges presented by the text: separation, ellipsis, segmentation of words or letter variation in Japanese.
5. The input must be raw text, either a single sentence or a full-length text.
6. The output should be a text annotated and parsed in XML, with the same tagset used in the annotation of the corpora.
7. The tool must also offer a recount of every type of modality and marker encountered.

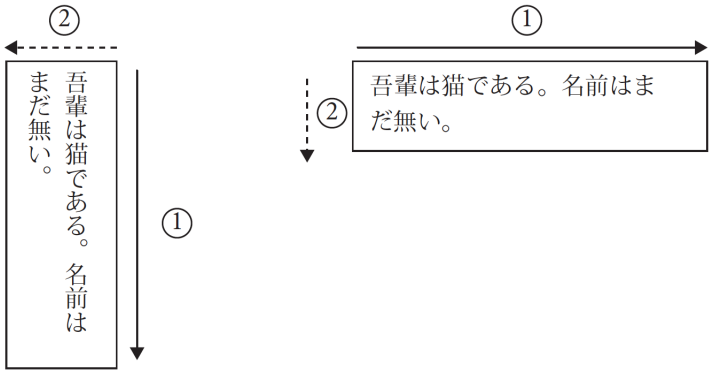
The basic idea is to continue working in both languages in a parallel way, to design a different script for each of them but that would function following the same steps. It works in three fundamental stages: tokenisation and POS tagging of the raw text, formation of a preliminary XML, and creation of the final XML structure. Figure 49 represents this design.

Figure 49: Design of the modality tagger



When processing a text, a program reads it according to its natural reading direction. Japanese can be written vertically and read from right to left (*tategaki*, 縦書き), but this style is normally limited to newspapers and novels (Obana, 1997) and rarely used in NLP tasks. The most common reading direction when processing is equal to Spanish, horizontal, and left to right –Figure 50, taken from Iwasaki (2013, p. 20).

Figure 50: Japanese traditional writing direction (left), Western style direction (right)



However, both languages are different in their syntactic ordering: Spanish is a SVO language, the auxiliary verbs precede the main verb, the predicates follow the copulative verb and the negation element normally precedes the negated item. Japanese is an SOV language, the auxiliaries and negation are attached after the

main verb. The main verbs, as well as the copulative verb, are located at the end of the sentence. Therefore, some differences may come in hand regarding the steps taken when processing each element. For example, in Spanish we find a negative element and then look for a modal marker situated under its scope, whereas in Japanese, we find a modal marker and then check its suffixes to see whether it is negated or not (compare Figures 51 and 52).

Figure 51: Order of processing in Spanish

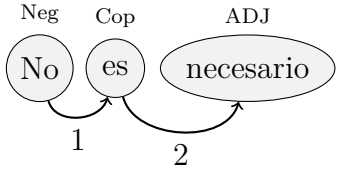
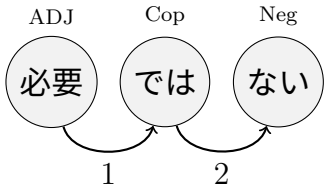


Figure 52: Order of processing in Japanese



5.1.1 Processing the raw text

Before directly looking for possible modal markers, the raw text is briefly prepared. First, it is tokenised manually into sentences, if formed by more than one. A selection of sentence separators was compiled, mainly full-stops, colons, semi-colons, and interrogation and exclamation marks, taking into account time stamps and abbreviations. It relies on ruled based tokenisation instead of machine learning although it might perform some errors (Tomanek et al., 2007).

After this, it is POS tagged by either Grampal in Spanish or Juman in Japanese. The purpose of this is to retrieve information that otherwise would be too time-consuming to retrieve manually, mostly the inflections of the verbs. In the case of Spanish, Grampal is used for the localisation of verbs in the subjunctive and imperative moods (See Table 98 for an example of the subjunctive).

Table 98: Example of Grampal's subjunctive tagging

Raw text	**Text tagged with Grampal**
No vayas tan rápido	No/NO/ADV **vayas/IR/V/sing,2,pres_subj** tan/TAN/Q rápido/RÁPIDO/ADJ/masc,sing

Juman, on the other hand, is used for a wider arrange of necessities, dealing with the typical problems of automatically processing Japanese texts. As commented earlier, the main issues are word segmentation and written variation. Juman would easily overcome these issues through the tokenisation and reading highlighting of each word. This would be very useful for words that could be written in the hiragana alphabet instead of kanji and vice-versa. For example, adverb 多分 (*tabun* 'probably') may be written also with the syllabic alphabet as たぶん. The first example in Table 99 shows the tagging example of a sentence with this adverb written in syllabic form. The most dangerous outcome of writing in hiragana are homonym words, which can be written and spelled in the same way but have a different meaning, kanji or POS tag. The combination of Juman's tokeniser, reading and POS tags will allow us to overcome these issues. Secondly, Juman is also used for automatically tagging potential and imperative moods, as Grampal. The second item of Table 99 includes an example of an imperative.

Finally, the Japanese POS tagger is used to retrieve the inflectional stem of each verb (See Table 6 in Chapter 2). In this initial stage of the program, each verb encountered will be annotated with an XML tag 'v' and the inflection will be included as an attribute. This will be useful to create the rules with the modal auxiliaries in the next step. The third item of Table 99 shows an example of these features. The verb 行きたい (*ikitai*, 'want to go') is formed with the auxiliary たい (*tai*) indicating desire, and the verb 行く (*iku*, 'to go') in its continuative form (基本連用形). Hence, in this stage the tagger will already mark the verb as <v type="Continuative"> 行き </v>.

Table 99: Example of Juman's tagging

Raw text	**Text tagged with Juman**
たぶん行きます。	たぶんたぶんたぶん副詞 8 * 0 * 0 * 0 " 代表表記: 多分/たぶん 修飾（ニ格" 行きいき行く動詞 2 * 0 子音動詞カ行促音便形 3 基本連用形 8 " 代表表記: 行く/いく付属動詞候補（タ系）ドメイン: 交通反義: 動詞: 帰る/かえる" ますますます接尾辞 14 動詞性接尾辞 7 動詞性接尾辞ます型 31 基本形 2 " 代表表記: ます/ます" 。。。特殊 1 句点 1 * 0 * 0 NIL
好きにしろよ。	好きにすきに好きだ形容詞 3 * 0 ナ形容詞 21 ダ列基本連用形 7 " 代表表記: 好きだ/すきだ反義: 形容詞: 嫌いだ/きらいだ動詞派生: 好く/すく" しろしろする接尾辞 14 動詞性接尾辞 7 サ変動詞 16 命令形 6 " 代表表記: する/する" よよよ助詞 9 終助詞 4 * 0 * 0 NIL 。。。特殊 1 句点 1 * 0 * 0 NIL
映画を見に行きたい。	映画えいが映画名詞 6 普通名詞 1 * 0 * 0 " 代表表記: 映画/えいがカテゴリ: 抽象物ドメイン: 文化・芸術" ををを助詞 9 格助詞 1 * 0 * 0 NIL 見み見る動詞 2 * 0 母音動詞 1 基本連用形 8 " 代表表記: 見る/みる補文ト自他動詞: 自: 見える/みえる" ににに助詞 9 格助詞 1 * 0 * 0 NIL 行きいき行く動詞 2 * 0 子音動詞カ行促音便形 3 基本連用形 8 " 代表表記: 行く/いく付属動詞候補（タ系）ドメイン: 交通反義: 動詞: 帰る/かえる" たいたいたい接尾辞 14 形容詞性述語接尾辞 5 イ形容詞アウオ段 18 基本形 2 " 代表表記: たい/たい" 。。。特殊 1 句点 1 * 0 * 0 NIL

5.1.2 Preliminary XML

After processing the text through the POS tagger, the program will create a preliminary XML with temporal tags of elements that may or may not be a modal marker or negative element. Particularly simple are markers that consist or rely only on one word: adverbs, adjectives and mood. Auxiliaries would require the program to check if the main verb, or the auxiliary in the case of Japanese, appears later in the text.

5.1.2.1 Negation

As we have seen in Section 4.6, negation of modal markers in Spanish is performed by negative elements –independent words such as adverbs and pronouns– preceding the marker by 2 words at most. In this preliminary stage the program will assign a temporal XML tag to all possible negative elements in the input text, checking them in a dictionary of different negative adverbs and pronouns extracted from the corpus and the literature. If they are followed by a formerly labelled 'stop' element, any of which that can break the scope of the negation, such as a comma or a pause, the negation annotation is discarded, as it will not affect the modal marker. In Japanese, negation is made by a negative suffix added at the end of the marker. Hence, after the modal marker is located, it will check if it is succeeded or not by it.

5.1.2.2 Adverbs

The Spanish tagger will check each word separated by Grampal in a dictionary of adverbs which contain the lemma and the probability value. Then, it assigns the corresponding XML tag. The same process is taken by the Japanese program, which also checks the reading tag by Juman. Table 100 shows an example of these dictionaries.

Table 100: Example of the tagger's adverb dictionaries

Adverb	Value
quizá	70%
quizás	70%
posiblemente	70%
probablemente	70%
inciertamente	50%
たぶん	70%
おそらく	70%
あるいは	50%
もしかすれば	50%

5.1.2.3 Mood

The imperative and potential moods are tagged using the POS taggers. When a subjunctive is encountered in a Spanish text, it checks if there is any negative element preceding it, with a distance of less than 3 words, as seen in the information recovered from the corpus (Table 88 from Section 4.6)

5.1.2.4 Adjectives

As explained in Chapter 2, the type of adjectives selected as modality markers are those considered predicative adjectives, or adjectives that serve as main element in sentences with a copula verb, more specifically, verbs *ser*, *estar* or *parecer*. The process is very similar to the one with negative elements: all the instances of these copulative verbs found using Grampal are marked with a temporal tag, as well as all adjectives found in the adjectives dictionary. Later on, if the copula and the adjective are in the same sentence at a distance inferior to 2 words, the temporal adjective is tagged as a definite modal marker.

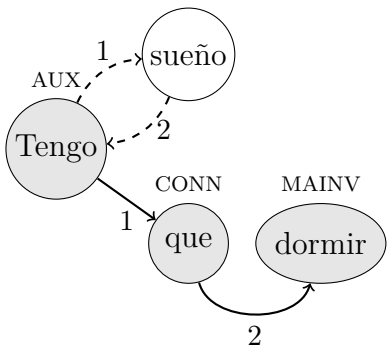
Japanese copula pertains a complicated issue, as it can be deleted at any time by the speaker (Chapter 2). For these elements, as with negation, the process is reversed due to the position of the verb at the end of a sentence. If a modal adjective candidate is found in the dictionary, it is annotated, and then the program will check if it is followed by the copula or, if omitted, if the adjective is situated at the end of a sentence.

5.1.2.5 Auxiliaries

Automatically locating modal auxiliaries involve additional steps in the processing of the text, as it requires finding the auxiliary verb and the main verb, if not omitted, and finding their negation.

Once the text is tagged using Grampal, the Spanish program will look for auxiliary verbs in the text that may form a modal periphrastic construction, and look for a following verb in infinitive form. If it is not followed by a main verb, a temporal tag is assigned in case there is a separation or omission of any of them, adding the 'id' tag (Figure 53).

Figure 53: Detecting Spanish modal periphrases



The same process is followed in Japanese, although the type of inflection of the main verb is taken into account. As explained in Chapter 2, Japanese stems subcategorise a specific auxiliary. That is, not all auxiliaries can be attached to every stem. A series of rules have been created that first checks the inflection of the verb, established by Juman and annotated in the previous step with an XML tag, and then looks ahead and checks if it is followed by the pertinent auxiliary. It is a hand-crafted rule based procedure inspired by similar studies such as Uchiyama et

al.'s detection of Japanese compound verbs (2005) and Murawaki and Kurohashi's detection of unknown morphemes (2008). Table 101 shows the compatibilities of stems and modal auxiliaries:

Table 101: Japanese inflection stems' (Juman) subcategorisation

Inflection	**Auxiliaries**
連用形 (Adverbial or Continuative)	*tai*, *kaneru*, *nakerebanaranai*, *kudasai*
基本形 (Plain)	*beki*, *shikanai*, *wakenihaikenai*, *shinobinai*, *hazu*, *tsumori*, *kotogadekiru*, *kamoshirenai*, *chigainai*, *kagiranai*, *oyobanai*
未然形 (Irrealis or Negative)	*nakerebanaranai*, *zaruwoenai*
条件形 (Conditional)	*(reba)ii*, *(tara)ii*
タ形 (ta-Form)	*hōgaii*, *shikanai*, *hazu*, *tsumori*, *kamoshirenai*, *chigainai*
テ形 (te-Form)	*temoii*, *moraitai*, *tehanakya*, *kudasai*, *hoshii*

In the preliminary XML the program will analyse the morphemes that follow these tags. If they are modal auxiliaries compatible with the type of inflection they follow, the complete construction (main verb + auxiliary) will be tagged as a modal marker. For example, to form the most frequent deontic necessity marker, なければならない (*nakerebanaranai*, 'have, must'), the main verb has to be in its irrealis form (See Table 6 in Chapter 2)[1]. Hence, if the auxiliary なければならない, or any of its possible variations, appears after a <v type="Irrealis"> tag, it will be marked as a modal marker.

If there is not a modal auxiliary after the verb, nothing is annotated, with the exception of できる (*dekiru* 'can, may') which, as we saw in the previous chapter, is the only auxiliary that may appear separated or alone with the main verb omitted (See Table 93). A temporal tag is assigned to it with the 'id' attribute.

In this stage, just like the types of writing systems used in the previous step, it

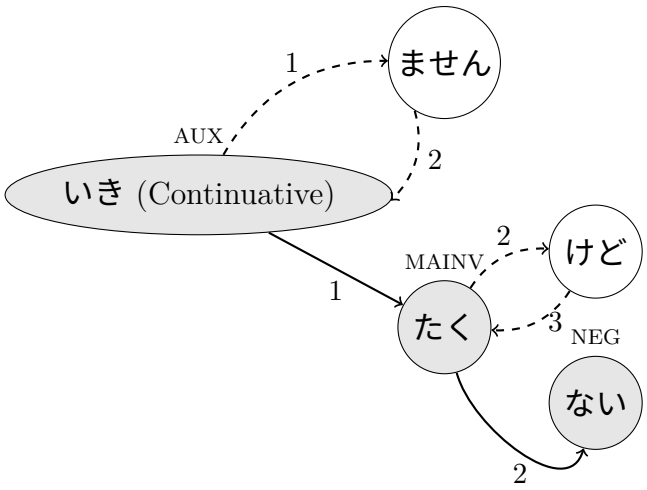
[1]Juman in some cases such as with the auxiliary する (*suru*, 'to do'), assigns the continuative stem tag.

is important to take into account the possible different variations that a Japanese marker may have. For example, the above mentioned なければならない may have all the following combinations in its present form depending on the degree of formality, style or strictness of the speaker, which can triple its size if considered negative and past forms of the proposition and past forms of the auxiliary:

- なければならない (*nakerebanaranai*)
- なければなりません (*nakerebanarimasen*)
- なければいけない (*nakerebaikenai*)
- なければいけません (*nakerebaikemasen*)
- なければだめ (*nakerebadame*)
- なけりゃならない (*nakeryanaranai*)
- なけりゃなりません (*nakeryanarimasen*)
- なけりゃいけない (*nakeryaikenai*)
- なけりゃいけません (*nakeryaikemasen*)
- なけりゃだめ (*nakeryadame*)
- なきゃならない (*nakyanaranai*)
- なきゃなりません (*nakyanarimasen*)
- なきゃいけない (*nakyaikenai*)
- なきゃいけません (*nakyaikemasen*)
- なきゃだめ (*nakyadame*)
- なけりゃ (*nakerya*)
- なきゃ (*nakya*)
- なくちゃ (*nakucha*)
- なくてはいけない (*nakutehaikenai*)
- なくてはいけません (*nakutehaikemasen*)
- なくてはだめ (*nakutehadame*)
- なくちゃいけない (*nakuchaikenai*)
- なくちゃいけません (*nakuchaikemasen*)
- なくちゃだめ (*nakuchadame*)

The program will consider every possible variation of the modal auxiliary as well as its negative or past tense suffixes, if any. In some occasions the variations may lead to the usage of another different inflection than the default one, such as the forms in the previous example starting with the negative *naku* followed by *te*-form and *ha* (i.e. *nakutehaikenai* and *nakutehadame*). As Figure 54 shows, detecting Japanaese auxiliaries is a very similar process to the Spanish counterpart:

Figure 54: Detecting Japanese modal auxiliaries



5.1.3 Final XML

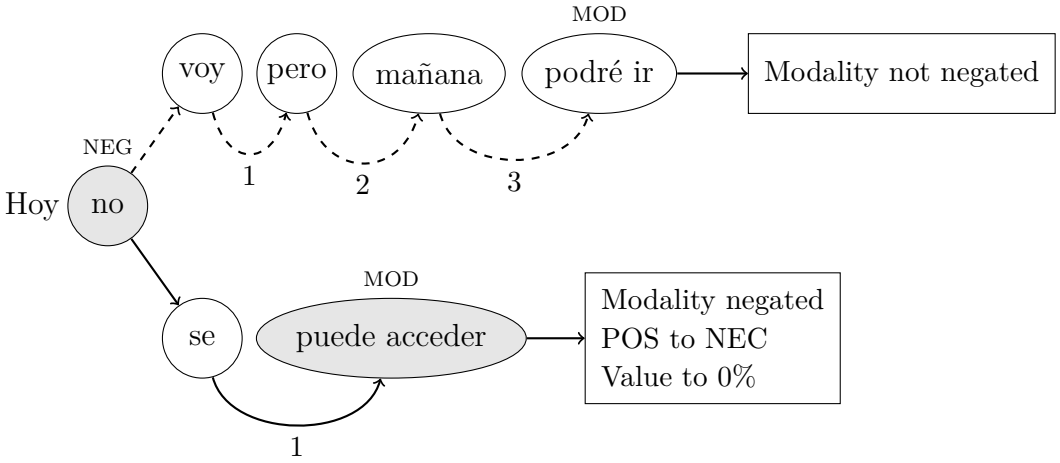
The last section of the modality tagger is the generation of the final XML that will serve as the output. Five main processes take place in this step for each language:

1. Checking negation (and modality modification), copulative and adjectives, separation of auxiliaries
2. Cleaning unnecessary and temporal tags
3. Counting modality
4. Checking XML syntax
5. Printing XML output and modality counts.

The tagger will now count the words between the auxiliaries and main verbs, adjectives and copula, and negation and modals in Spanish. For example, with

negation, if the count from the modal marker is less than 2 words, it will consider the marker as negated. Then, it will change the attribute 'neg' to 'yes' and modality type from 'Possibility' to 'Necessity' or vice-versa if the modal marker belongs to the type that changes with negation, specified in the dictionary and rules. The value is set to 0% if it has changed to necessity from possibility and to 50% if otherwise. If the modal is not changed with negation, the probability value remains unchanged. The Japanese part will follow the same steps although the negative suffix was already checked in the previous step. At this stage, the program will deal only with changing the type of Japanese modality. Figure 55 shows an example of this process.

Figure 55: Detecting negation



Similar steps will be taken with the copula and the modal adjectives candidates, and the possible separated auxiliaries. If the Spanish main verb or the Japanese auxiliary is found several words after, the corresponding 'id' and 'ref' attributes are added to the modality tag. If not, the auxiliary will be left as a possible modal auxiliary construction with an elliptic main verb, and the attribute 'elli' is added.

After this, temporal tags are stripped and cleaned from the XML, the syntax is checked and the document is created, and the modality is counted according to its first and second levels, and the grammatical class of the marker. The following tables (102 and 103) contain some modal marker examples extracted from the corpora and the results of each step of the program.

Table 102: Processing steps of several Spanish examples

Input text	**Processed text**	**Preliminary XML**	**Output XML**
Quizás lo retrasen un poco.	<change /><m modtype="POSS" subtype="EPIS" class="Adverb" neg="no" value="70%">Quizás </m> lo retrasen un poco .	<s><change /><m modtype="POSS" subtype="EPIS" class="Adverb" neg="no" value="70%"> Quizás</m> lo retrasen un poco.</s>	<s><m class="Adverb" modtype="POSS" neg="no" subtype="EPIS" value="70%"> Quizás</m> lo retrasen un poco.</s>
Entonces tampoco era al final no fue necesario o sea que.	entonces tampoco <cop class="verb_BE" neg="no">era</cop> a el final <w neg="yes">no</w> <cop class= "verb_BE" neg="no">fue</cop> <adj modtype="NEC" subtype="EPIS" class="Adjective" neg="no" value="100%"> necesario</adj> o sea que .	<s>entonces tampoco <cop class="verb_BE" neg="no">era</cop> a el final <w neg="yes">no</w> <cop class="verb_BE" neg="no"><candid modtype="NEC" subtype="DEON" class="AUX" neg="no" value="100%">fue </candid></cop> <adj modtype="NEC" subtype="EPIS" class="Adjective" neg="no" value="100%">nece- sario</adj> o sea que.</s>	entonces tampoco era a el final <w neg="yes">no</w> fue <m class="Adjective" modtype="NEC" neg="no" subtype="EPIS" value="100%"> necesario</m> o sea que.</s>
Y también por eso porque en mi casa no podía estudiar, ¿sabes?	Y también por eso porque en mi casa <w neg="yes"> no</w> podía estudiar <stop>,</stop> ¿ sabes ?	<s>Y también por eso porque en mi casa <w neg="yes">no</w> <change /><m modtype="POSS" subtype="AMBG" class="AUX" neg="no" value="50%">podía estudiar</m> <stop>, </stop> ¿sabes?</s>	<s>Y también por eso porque en mi casa <w neg="yes">no</w> <m class="AUX" modtype="NEC" neg="yes" subtype="AMBG" value="0%">podía estudiar</m>, ¿sabes?</s>

Table 103: Processing steps of several Japanese examples

Input text	Processed text	Preliminary XML	Output XML
やっぱり最近の傾向なのかもしれませんね。	やっぱり最近の傾向なかもしれません <w type="ENDPART"> ね </w>。	<s> やっぱり最近の傾向な <m class="AUX" modtype="POSS" neg="no" subtype="EPIS" value="70%"> かもしれません </m><w type="ENDPART"> ね </w>。</s>	<s> やっぱり最近の傾向な <m class="AUX" modtype="POSS" neg="no" subtype="EPIS" value="70%"> かもしれません </m> ね。</s>
結構見られない、	結構 <v type="irrealis"> 見 </v><v type="potential"> られ </v> ない、	<s> 結構 <m modtype="NEC" subtype="DEON" class="mood_POT" neg="yes" value="0%"> 見られ <w type="neg"> ない </w></m>、</s>	<s> 結構 <m class="mood_POT" modtype="NEC" neg="yes" subtype="DEON" value="0%"> 見られない </m>、</s>
必要じゃないよ！	<w modtype="NEC" value="100%" subtype="EPIS" class="Adjective" neg="no"> 必要じゃ </w> ない <w type="ENDPART"> よ </w>。	<s><changem modtype="POSS" value="100%" subtype="EPIS" class="Adjective" neg="yes"> 必要 <w type="neg"> じゃない </w></changem> <w type="ENDPART"> よ </w>。</s>	<s><m class="Adjective" modtype="POSS" neg="yes" subtype="EPIS" value="50%"> 必要じゃない </m> よ。</s>
なんか年齢登録しなきゃいけないんでとかいってメール来て、	なんか年齢 <v type="surunoun"> 登録 </v><v type="continuative"> し </v> なきゃいけないんでとかいってメール <v type="te"> 来て </v>、	<s> なんか年齢 <v type="surunoun"> 登録 </v><m modtype="NEC" subtype="DEON" class="AUX" neg="no" value="100%"> しなきゃいけ <w type="neg"> ない </w></m> んでとかいってメール <v type="te"> 来て </v>、</s>	<s> なんか年齢登録 <m class="AUX" modtype="NEC" neg="no" subtype="DEON" value="100%"> しなきゃいけない </m> んでとかいってメール来て、</s>

The only process that could not be formalised into rules and was left out of the program was the annotation of Spanish errors. Much like ambiguity, it will certainly reduce in the epistemic/deontic attribute of the annotation. The most problematic would be the distinction between *Deber* (deontic) and *Deber de* as seen earlier, as the amount of mistakes made by natives is nearly 40%.

And, to conclude the chapter and the study, Figures 56 and 57 summarise the process of the program in each language with an example sentence.

Figure 56: Design of the Spanish modality tagger

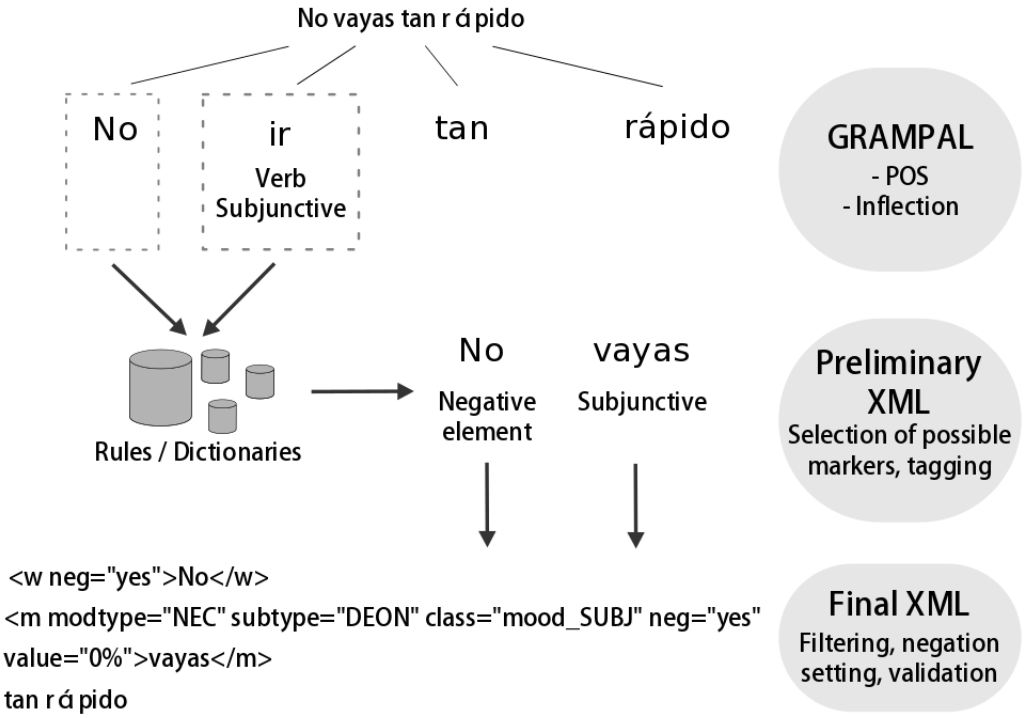
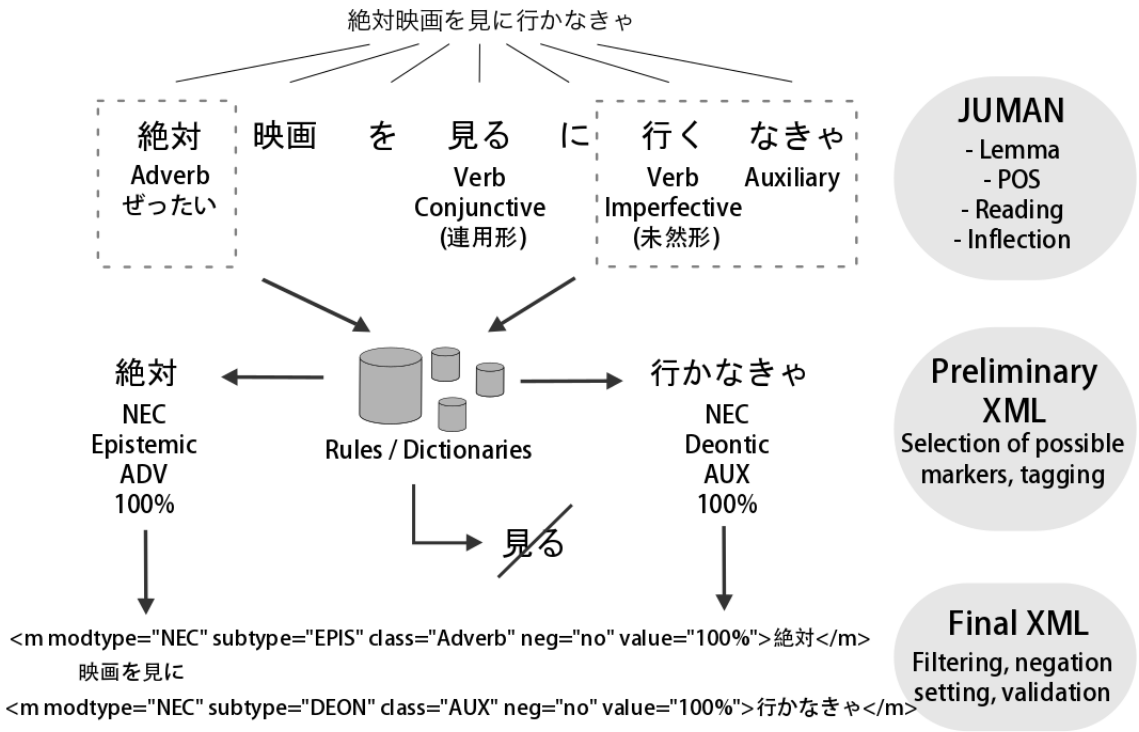


Figure 57: Design of the Japanese modality tagger



Chapter 6

Conclusions

6.1 Summary and conclusions

Chapter 2 examined the field of modality, observing its origins and major breakthroughs in its history that have led to today's situation of the area in European and Japanese linguistics. We can group the most widespread definitions of modality into three approximations: (1) all the elements that modify the proposition; (2) the expression of the attitude or subjectivity of the speaker; and (3) the relation between language and truthfulness of reality. We have decided that the first two are not suitable for a study with a computational, cross-linguistic idea in mind. To annotate everything outside a proposition seems out of proportion and does not provide concrete semantic information from the text (Is a Spanish discourse marker at the same level as a Japanese case particle?) Similarly, to study the subjectivity of the speaker in a text seems more appropriate for a study closer to the Pragmatics field and almost impossible to formalise into hand-written rules for a tagger. We believe the third option, to consider modality a semantic element encoded in a part of the sentence such as aspect or tense, is much more approachable for this kind of study.

Here, modality signals how close is a state of affairs (SOA) to becoming true, either by a belief or a desire of the speaker. If the SOA is or is not true in every possible world, it is considered a necessity. If, on the other hand, the SOA may be true, it is considered a possibility. If the truth condition is triggered by the belief of the speaker it is an epistemic modality, and if it is realised by a desire, it is deontic.

Therefore, as tense and aspect, it is a semantic category realised by certain elements in the sentence. We have selected those that overtly mark it and complement the indicative and subjunctive moods: auxiliaries and suffixes attached to a verb, adverbs, predicative adjectives fulfilling the meaning of the copula, and the imperative and potential moods.

Chapter 3 established the methodology and steps taken in the study, and described the tools used and made clear the tagset used for annotating the corpora and the tagger. It also provided a list and description for all possible modal markers as well as their corresponding tag.

We can already see in this stage differences between both languages, especially in the auxiliaries, which are the most common way of encoding modality. Spanish uses 7 modal auxiliaries, and only one of them is used for possibility when not negated. There are 23 auxiliaries in Japanese, and only 4 of them are used for possibility uses. Necessity is a much more complex value, especially in Japanese. Also, in both languages, like in English, German or Italian, negation does not affect each marker equally. Some of them change into the opposite type of modality (the auxiliary is negated), other maintain it (when the main verb is negated).

Regarding the remaining markers, both languages also have a series of adverbs and predicative adjectives indicating modality, but Spanish offers a wider array of elements than Japanese. Also, the imperative is used exclusively as a modal element, in addition to the potential in Japanese.

Chapter 4 is dedicated to the quantitative study performed on the data extracted from both corpora. We have concluded that modality is present in both languages, but it is highly used and more irregular on average in Spanish. Japanese frequencies are lower, but much more constrained and similar among speakers. However, this difference is located primarily in the usage of necessity markers. Speakers from both languages use nearly the same amount of possibility modality. With necessity, Japanese is much more constrained than Spanish, especially when two or more speakers engage in a direct conversation, the only situation where the amount of the necessity is nearly equal to possibility.

The epistemic and deontic comparison has shown an extremely high amount of ambiguity in Spanish. This makes it unreliable for automatic processing. Qualitative studies, such as the ones located in the pragmatics or cognitive linguistics area may find it attractive to disentangle this ambiguity looking at the underlying intention or mental process of the speaker. For a quantitative study, a necessity/possibility distinction seems much more attractive and clear, although the information retrieved is more general.

On the other hand, the grammatical class of the marker hardly changes according to the type of discourse: auxiliaries are by far the most common way of signalling

modality, followed by imperative mood in Spanish and adverbs in Japanese.

The comparisons according to non-linguistic elements have indicated that the usage of modality is similar depending on the gender and age of the speakers, with a slightly preference for necessity in women, deontic in Spanish women and epistemic in Japanese and adjectives over mood. Overall, it appears modality is a phenomena much more related to the type of interaction between humans and social restrains than factors such as gender and age.

The chapter finishes with an analysis of elements in the text that can modify the markers. The most important conclusion drawn is that negation is highly frequent, almost equally in both languages, and a feature that must be addressed by the automatic annotation. Separation and ellipsis is possible but not significantly frequent.

The last chapter, Chapter 5 described the development of the modality tagger for both languages. The program has been built upon what was learnt from both the theoretical and empirical information described in the previous chapters. It has been designed with hand-crafted rules following the same procedure and tagset for both languages. (1) Pre-processing of the raw text: sentence tokenisation and POS tagging; (2) Preliminary XML: annotation of potential modal markers, negative elements, and verbal information. (3) Final XML: processing changes in modality by negation, separation and ellipsis. The program has been made available online in the UAM Computational Linguistics Laboratory under the address `http://elvira.lllf.uam.es/modtag/mainmodtagger.html`, and is intended to be used in future studies with a wider array of texts.

6.2 Final remarks

A sense of desire for knowledge has been the main motivation of this work, a desire to observe differences and similarities in two apparently very different languages. The most attractive characteristic of modality is its universality and its relation to human reasoning. More than comparing language elements independently we wanted to observe if the way of using language is similar between speakers through patterns of use. The way of carrying out the study has been through a computer, handling large amounts of text and data and aiming to automatically find these patterns more quickly for replicating studies in the future with new data.

As discussed above, there are two basic procedures when developing an automatic way of extracting information from data: either by developing hand-written rules, or by letting the computer decide through machine learning. Neither way is better than the other, the decision depends on the objective and nature of the work. For this one the first option was taken. Machine learning is an extremely powerful tool and achieves optimal results. However, its downfall is that it is limited to a series of probability calculations. This may work well for developing a commercial product, but the linguistic information, the most interesting element for a linguist, is lost in a 'black box'. Hand-written rules may be more limited and achieve poorer results in recall indexes, but are closer to theoretical implications of language and let the linguist stay in control and learn how language works. We may need to apply machine learning operations in the future to improve the tagger, such as for example, speakers' errors, or, if handled carefully, the ambiguity issue, but we wanted its foundations to be controlled and reinforced by theoretical and empirical information. Theory is useless if not observed how it is present in *real* language, and empirical data is uncontrolled without some underlying theoretical implications.

This work may not have immediate applications. However, we do believe it is a first step, using quantitative and computational methords, to bring Spanish and Japanese closer together. Although countless work has been done comparing Spanish and Japanese modality, it remains on an abstract theoretical level, without using real texts made by native speakers and statistical information, which is essential. The

development of a program that can automatically find modal markers can allow us to repeat the study more quickly in a variety of different texts and discourses, from spoken to written, and specialised ones like media and financial texts, to expand our knowledge in these languages and use it in the future for any desired purpose, from teaching to translation.

6.3 Limitations and future work

The study has been designed with a general idea in mind: simplicity. The definition of modality should be as clear and concrete as possible to fit both languages and hand-crafted annotation rules. Nevertheless, this approximation has a series of limitations or downfalls as well as additional improvements that should be taken into account in the future:

- Broaden the number of modal markers.
- Consider the relation between modality and tense.
- Include adverbs that can modify the probability value of markers.
- Extend the study to new texts and corpora.
- Perform an evaluation of the modality tagger.

Many possible markers may have been left out. We have selected only those *marked* elements, but necessity and possibility is also present in other parts of the sentence. The clearest one is the indicative mood, that can imply an absolute certainty of an event (necessity) such as an intention (deontic) or factuality (epistemic) or even orders (deontic). As another example, for imperative actions we have only considered the imperative mood and a series of auxiliaries, but in Japanese these forms may seem too direct and threatening. Speakers could prefer the present indicative form for this purpose, the hortative -*yō* or the -*te* form of the verb, an informal simplification of the marker -*tekudasai* (てください). We insist on creating a clear line of what is considered and marked as modality, even though it may leave out elements that convey similar meanings in the sentences. However, the line can be expanded in future studies. For example, we can consider modality to be located in the semantic information of a series of lexical verbs such as *necesitar* ('need') or *querer* ('want') in Spanish, or the previously mentioned -*te* form of Japanese verbs. A new study considering these forms may show us a very different picture to the one presented in these pages.

Following this, an aspect that has not been tackled in these pages is the relation between modality and tense. Both are common elements of human languages, and

certainly are related to each other and have been studied before by logicians and linguists, but we have not explained how this relation can take place and how it is used in the corpora. Should an auxiliary in the past tense always indicate necessity as an event that has already occurred would be, by definition, true? Spanish, in opposition to Japanese, can overtly mark future tense with verb inflection including, in some cases, a sense of intention by the speaker, similarly to the modal auxiliary *ir a*. If both elements are so similar, should we include the future tense suffix as a modal marker? Also, the same can be represented in Spanish and Japanese with the combination future temporal adverbs and the plain indicative form, so these cases could also be treated as necessity modality. This could be addressed in future studies using the same corpora.

Another feature left out in these pages but that could also be studied in the future is the way different partially-negative or quantitative adverbs can increase or decrease the probability value. Similar to negation, the value of a possibility marker can be slightly modified. For example, the Spanish adjective *probable* has a value assigned of 70%. However, if adverbs *poco* ('few') or *muy* ('very') are assigned to it, the probability should move to around 30% and 90% respectively. A higher degree of modal meaning complexity can be added to the tagger in future studies.

The development of the tagger can allow us to repeat the quantitative analysis more rapidly in new texts and corpora. The study could be repeated in additional spoken texts to provide further evidence on the usage of modality in spontaneous speech, or in new types of discourse and register for comparison. Especially for Japanese, to see if the positive outcome of the normality tests is maintained. Also, since this study has been developed with a dependency syntax in mind it would be interesting to improve the tagger to work with dependency treebanks, adding modal semantic information to the appropriate nodes of the syntactic tree.

Finally, the tagger needs an evaluation to formally check its performance. A solid evaluation needs to be performed, probably using different types of texts to verify if the rules, created from the observation in spoken corpora, are still valid. This would require an amount of time and workforce that extends the duration and budget of this study, but which definitely needs to be made in the near future.

Chapter 7

Conclusiones (Spanish)

7.1 Resumen y conclusiones

En el Capítulo 2 hemos estudiado el concepto de modalidad y observado sus orígenes y mayores avances en su historia que han llevado a la situación actual en la lingüística europea y japonesa. Podemos agrupar las interpretaciones más extendidas de la modalidad en tres aproximaciones: (1) la modalidad está formada por todos aquellos elementos que pueden modificar una proposición; (2) la modalidad es la expresión de la actitud o subjetividad del hablante; (3) la modalidad es la relación entre el lenguaje y el nivel de verdad en la realidad. Hemos decidido que las dos primeras definiciones no son apropiadas para un estudio lingüístico comparativo y computacional. Anotar todo aquello fuera de una proposición parece desproporcionado y no nos otorga un contenido semántico concreto (¿está un marcador discursivo español al mismo nivel que una partícula de caso japonesa?). De forma parecida, estudiar la subjetividad del hablante en un texto parece más apropiado para un estudio pragmático y resulta prácticamente imposible formalizar en reglas para un etiquetador automático. Creemos que la tercera opción, considerar la modalidad como un elemento semántico de la oración al igual que el tiempo o el aspecto, es mucho más apropiada para la naturaleza del estudio.

En nuestro caso, la modalidad indica cómo de cerca está una situación de convertirse en verdadera, a través una creencia o un deseo del hablante. Si la situación es o no verdadera en todos los mundos posibles, se considera una necesidad. Si, por otro lado, la situación es verdadera en solo alguno de los mundos, estaremos hablando de una posibilidad. Si la condición de verdadero está movida por una creencia del hablante, la consideramos una necesidad o posibilidad epistémica, mientras que si es por un deseo, la llamaremos deóntica.

Por tanto, de la misma manera que el tiempo y aspecto, la modalidad es una categoría semántica que está presente en todas las lenguas y se representa en la oración por una serie de elementos llamados marcadores modales. Hemos seleccionado para este estudio aquellos marcadores que codifican la modalidad de forma una marcada y abierta, y que complementan los modos indicativo y subjuntivo: auxiliares, sufijos, adverbios y adjetivos predicativos que amplían el contenido semántico de la cópula,

así como los modos imperativo y potencial.

En el Capítulo 3 hemos establecido la metodología y los pasos seguidos en este estudio, descrito las herramientas utilizadas, y explicado el listado de etiquetas usadas para anotar la modalidad en los corpus y por el etiquetador automático. Hemos incluido también un listado y descripción de cada marcador modal posible, así como su etiqueta correspondiente.

En este punto del estudio podemos empezar a ver características significativas en ambas lenguas, especialmente en los auxiliares, el elemento más común para marcar la modalidad. El español utiliza 7 auxiliares modales distintos, y solamente uno indica posibilidad. En japonés hay 23 auxiliares, y solamente 4 de ellos tienen usos de posibilidad. La necesidad es un valor altamente complejo, especialmente en japonés. Además, en ambos idiomas, como en inglés, alemán o italiano, la negación no afecta de la misma manera a cada marcador. Algunos de ellos cambiarán al tipo contrario de modalidad cuando el auxiliar recibe la negación, mientras que otros lo mantienen cuando es el verbo principal el que está negado.

El Capítulo 4 está dedicado al estudio cuantitativo realizado en los datos extraídos de ambos corpus. Hemos concluido que la modalidad está presente extensivamente en ambas lenguas, aunque tiene una frecuencia mayor y más irregular en español. Los números en japonés son menores, pero están más concentrados y son más regulares. Sin embargo, esta diferencia se sitúa principalmente en la necesidad. Los hablantes de ambas lenguas usan prácticamente la misma cantidad de marcadores de posibilidad; mientras que con la necesidad el japonés está mucho más restringido que el español, especialmente en los diálogos, cuando hay interacción directa entre los hablantes, el único momento en el que la cantidad de necesidad es casi idéntica a la de posibilidad.

La comparación entre marcadores epistémicos y deónticos muestra una ambigüedad muy elevada en español. Esto la convierte en algo poco fiable para el procesamiento automático. Estudios cualitativos en pragmática o lingüística cognitiva pueden encontrar atractivo separar y estudiar este solapamiento de significados, observando la intención o proceso mental llevado a cabo en el hablante. Sin em-

bargo, un estudio cuantitativo automático debería centrarse en la distinción necesidad/posibilidad, a pesar de contener unos significados más generales.

La categoría gramatical de los marcadores apenas cambian con el tipo de discurso: los auxiliares son el elemento más común de marcar modalidad, seguidos por el modo imperativo en español y los adverbios en japonés.

Las comparaciones respecto a variables no lingüísticas han mostrado que el uso de la modalidad es prácticamente el mismo en cuanto a sexo y edad de los hablantes, con una ligera preferencia por la necesidad en las mujeres, marcadores deónticos en las mujeres españolas, y epistémicos en las japonesas, así como un mayor uso de adjetivos en lugar de modo gramatical. Por lo general parece que la modalidad es un fenómeno más relacionado con el tipo de interacción entre personas y restricciones sociales que factores como el género o la edad.

El capítulo termina con un análisis de los elementos en el discurso que pueden modificar los marcadores modales. La conclusión más importante extraída es la alta frecuencia de la negación, casi la misma en ambas lenguas, que obliga a ser procesada por el etiquetador automático. Separación y elipsis son también posibles aunque no con tanta frecuencia.

En el último Capítulo 5 hemos descrito el desarrollo del etiquetador automático de la modalidad para ambas lenguas. El programa ha sido construido basándose en lo aprendido del ámbito teórico y la información empírica extraída de los corpus, descrita en los anteriores capítulos. Se ha diseñado con reglas escritas a mano siguiendo el mismo procedimiento y listado de etiquetas para ambos idiomas: preprocesamiento del texto de entrada: el texto se divide por oraciones y se etiqueta morfológicamente; XML preliminar: búsqueda de elementos de negación, información verbal y candidatos a marcadores modales; generación de XML final de salida: procesamiento de cambios causados por la negación, separación y elipsis. El programa está disponible en la página web del Laboratorio de Lingüística Computacional de la UAM en la dirección `http://elvira.lllf.uam.es/modtag/mainmodtagger.html`, y se pretende usar en el futuro para realizar estudios más rápidamente en textos más amplios y de distinta naturaleza.

7.2 Apuntes finales

La mayor motivación de este trabajo ha sido saciar el deseo de descubrir diferencias y similitudes en dos idiomas aparentemente muy distintos. El aspecto más atractivo de la modalidad es su universalidad y su relación con el razonamiento humano. Más que comparar solamente elementos lingüísticos, hemos querido estudiar cómo son usados por los hablantes a través de patrones de uso, y observar si ese uso es significativamente diferente. La manera de llevar a cabo este estudio ha sido a través de un ordenador, usando grandes cantidades de texto y datos con el objetivo de encontrar estos patrones y reproducirlos rápidamente en estudios futuros.

Como hemos comentado antes, existen dos procedimientos básicos para extraer información automáticamente de textos. Podemos escribir reglas a mano, o dejar un ordenador hacerlo a través de aprendizaje automático. Ninguna de las dos maneras actúa mejor que otra, la decisión depende de la naturaleza del trabajo. Para esta tesis hemos tomado la primera opción. Aprendizaje automático es una herramienta extremadamente potente que obtiene resultados óptimos. Sin embargo, está limitada a una serie de cálculos de probabilidad. Puede funcionar bien para desarrollar un producto comercial, pero la información lingüística, lo más valioso para un lingüista, se pierde en una caja negra. Desarrollar reglas a mano puede estar más limitado y obtener resultados peores en índices de precisión, pero están más cerca de las implicaciones teóricas del lenguaje y permiten al lingüista mantener el control y aprender cómo funciona una lengua. Probablemente necesitemos aplicar aprendizaje automático en el futuro para aumentar la cobertura del etiquetador para resolver el problema de los errores en los hablantes nativos, por ejemplo, o incluso la ambigüedad. Sin embargo, hemos preferido que las bases del programa estén controladas y reforzadas por conocimiento teórico y empírico, ambos necesarios. La teoría es inútil si no es observada en lenguaje *real*, y los datos empíricos están descontrolados sin estar sustentados en reglas teóricas.

Este trabajo puede no tener aplicaciones inmediatas, pero sí creemos que es un primer paso a la hora de acercar el español y el japonés, un área en la investigación contrastiva que casi no ha recibido atención por parte de estudios cuantitativos y

computacionales. Se han escrito un sinfín de trabajos comparando ambas lenguas, pero siempre terminan en un nivel abstracto, sin usar textos reales producidos por hablantes nativos, que resultan esenciales. A través del desarrollo de un etiquetador automático de la modalidad pretendemos repetir el estudio en una variedad más rica y amplia de textos y tipos de discurso, por ejemplo observando el uso en textos escritos, o especializados como en los campos periodísticos o financieros, y así expandir nuestro conocimiento de estas lenguas para poder usarlo en el futuro en campos como la enseñanza o la traducción.

7.3 Limitaciones y trabajo futuro

Este estudio se ha llevado a cabo con una idea general en mente: simplicidad. La definición de la modalidad tiene que ser lo más clara posible para encajar en ambas lenguas y ser formalizado en reglas de anotación. Sin embargo, esta aproximación cuenta con una serie de limitaciones e inconvenientes que tienen que ser tratadas y mejoradas en el futuro:

- Ampliar el número de marcadores modales.
- Estudiar la relación entre modalidad y tiempo.
- Incluir adverbios que puedan modificar el valor de probabilidad de los marcadores.
- Extender el estudio a nuevos textos y corpus.
- Realizar una evaluación del etiquetador de modalidad.

Un gran número de marcadores pueden haberse quedado fuera del estudio. Solamente se han seleccionado aquellos elementos *marcados*, pero la necesidad y la posibilidad puede estar presente en otras partes de la oración. El más claro es el modo de indicativo, que puede indicar una certeza absoluta de un evento (necesidad), como una intención (deóntica), un hecho (epistémica) o incluso órdenes (deóntica). Otro ejemplo son las construcciones que pueden indicar una acción imperativa además de los auxiliares y modos que hemos tratado en este estudio. En japonés estas formas pueden resultar demasiado directas y los hablantes pueden preferir maneras más indirectas, como la forma indicativa presente, la terminación hortativa *-yō*, o la forma *-te* de un verbo, una simplificación informal del modal *-tekudasai* (てください). Para un primer estudio insistimos en crear una línea clara delimitando aquello que consideramos modalidad, incluso si eso conlleva dejar candidatos potenciales de lado. Sin embargo, esta línea puede ser ampliada en estudios futuros. Otro ejemplo a tener en cuenta son aquellos verbos que por su contenido semántico pueden indicar algún valor modal, como los españoles *necesitar* o *querer*. Un nuevo estudio que tenga en cuanta formas adicionales puede darnos información más extensa de la que ha sido presentada en estas páginas.

Continuando con el trabajo futuro, un aspecto que no ha sido tratado en este estudio es la relación entre modalidad y tiempo. Ambos son elementos comunes en el lenguaje humano, y sin duda existe una relación entre ellas que ha sido estudiada con anterioridad por lógicos y lingüistas, pero no hemos explicado esta relación ni cómo se presenta en el corpus. ¿Un auxiliar en tiempo pasado debería indicar siempre una necesidad, puesto que marca un evento que ya ha ocurrido y es, por tanto, verdadero? Además, el español, al contrario que el japonés, puede marcar abiertamente el tiempo futuro a través de la flexión verbal, incluyendo en algunas ocasiones una intención por parte del hablante, de forma parecida al modal *ir a*. Si ambos elementos son tan similares, ¿deberíamos incluir el tiempo futuro como marcador modal? Incluso, el mismo significado se puede conseguir combinando adverbios temporales futuros y el tiempo presente tanto en español como en japonés, sugiriendo también la posibilidad de incluir estos casos como una marcación de la necesidad. Esto puede ser tratado en el futuro usando incluso los mismos corpus.

Otro elemento que podemos estudiar en el futuro es la manera en la que adverbios parcialmente negativos o cuantitativos pueden aumentar o descender el valor de probabilidad. Dicho de otra manera, de forma parecida a la negación, el valor de una posibilidad puede verse modificado por un adverbio. Por ejemplo, el adjetivo modal español *probable* tiene asignado un valor del 70%. Sin embargo, si se modifica con los adverbios *poco* o *mucho*, el valor debería moverse a un 30% o un 90%, respectivamente. Podemos ampliar la complejidad de la clasificación realizada por el etiquetador automático en estudios futuros.

Como hemos comentado anteriormente, el etiquetador nos permite repetir el análisis cuantitativo más rápidamente en textos y corpus nuevos. Podemos repetir el estudio en textos orales adicionales para extraer una evidencia más sólida sobre el uso de la modalidad en el discurso espontáneo, o usar distintos tipos de textos y registros para hacer estudios comparativos. Particularmente atractivo sería comprobar en nuevos textos orales si se mantiene la distribución normalizada obtenida en nuestro corpus japonés, si realmente es la modalidad un fenómeno tan regular en este idioma, o si por el contrario los resultados son debidos al reducido tamaño del corpus. Además, al haber diseñado este estudio con la sintaxis de dependencias

en mente, puede ser interesante expandir el etiquetador usando *parsers* sintácticos para el español y el japonés con el fin de generar árboles de dependencias incluyendo información semántica modal.

Finalmente, el desarrollo del etiquetador automático requiere una evaluación apropiada para comprobar objetivamente su rendimiento. Desarrollar una evaluación sólida requiere una cantidad de tiempo y trabajo que se escapa de los límites de este trabajo, pero tiene que ser realizada definitivamente en el futuro. Podría realizarse, por ejemplo, utilizando distintos tipos de textos y comprobar si las reglas creadas son igualmente válidas.

References

Abney, S. (2011). Data-Intensive Experimental Linguistics. *Linguistic Issues in Language Technology*, *6*(2), 1–27.

Aikhenvald, A. Y. (2005). *Evidentiality*. Oxford University Press, USA.

Akiba, D. (2014). Interpreting modals by phase leads. In E. Leiss & W. Abraham (Eds.), *Modes of Modality: Modality, typology and Universal Grammar.* Amsterdam: John Benjamins Publishing.

Asahara, M., Yoneda, R., Yamashita, A., Den, Y., & Matsumoto, Y. (2002). Use of XML and relational databases for consistent development and maintenance of lexicons and annotated corpora. In *Proceedings of the Third International Conference onf Language Resources and Evaluation (LREC 2002)* (pp. 1372–1378). Las Palmas, Spain.

Austin, J. L. (1975). *How to Do Things with Words: Second Edition (The William James Lectures)*. Cambridge: Harvard University Press.

Baker, K., Bloodgood, M., Dorr, B. J., Callison-Burch, C., Filardo, N. W., Piatko, C., … Miller, S. (2015). Use of Modality and Negation in Semantically-Informed Syntactic MT. *Computational Linguistics*, *38*(2), 411–438.

Bally, C. (1950). *Linguistique generale et linguistique francaise [General Linguistics and French Linguistics]*. Bern: Francke.

Baroni, M., & Evert, S. (2008). Statistical methods for corpus exploitation. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. an international handbook* (pp. 777–803). Berlin: Mouton de Gruyter.

Biber, D. (1991). *Variation across speech and writing.* Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* London: Longman.

Bosque, I. (1980). *Sobre la Negación.* Madrid: Cátedra.

Bosque, I. (2012). Mood: Indicative vs. Subjunctive. In J. I. Hualde, A. Olarrea, & E. O'Rourke (Eds.), *The Handbook of Hispanic Linguistics.* Chichester: Wiley-Blackwell.

Briz Gómez, A. (2001). *El Español Coloquial en la Conversación. Esbozo de Pragmagramática. [Colloquial Spanish in the Conversation. A pragmagramatic outline.].* Barcelona: Ariel.

Brown, P., Levinson, S., & Gumperz, J. (1987). *Politeness: Some universals in language usage* (Reissue ed.). Cambridge: Cambridge University Press.

Bybee, J. (1985). *Morphology: A Study of the Relation Between Meaning and Form* (1st ed.). Chicago/London: John Benjamins Publishing Company.

Bybee, J., Perkins, R., & Pagliuca, W. (1994). *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World.* Chicago: University Of Chicago Press.

Camacho, J. (2012). Ser and Estar: The individual/stage-level distinction and aspectual predication. In J. I. Hualde, A. Olarrea, & E. O'Rourke (Eds.), *The Handbook of Hispanic Linguistics* (pp. 453–476). Chichester: Wiley-Blackwell.

Cepeda, G., & Poblete, M. T. (2006). Cortesía verbal y modalidad: Los marcadores discursivos. *Revista Signos*, *62*(39), 357–377.

Clinque, G. (2006). *Restructuring and functional heads. The cartography of syntactic structures.* Oxford: Oxford University Press.

Corder, S. P. (1967). The significance of learner's errors. *International Review of Applied Linguistics in Language Teaching*, *5*(4), 161–170.

Cornillie, B. (2007). *Epistemic Modality and Evidentiality in Spanish (Semi) Auxiliaries. A Cognitive-functional Approach.* Berlin/New York: Mouton de Gruyter.

Cornillie, B. (2009). Evidentiality and epistemic modality: On the close relationship of two different categories. *Functions of Language*, *16*(1), 44–62.

Cornillie, B. (2010). An interactional approach to evidential and epistemic adverbs in Spanish conversation. In *The Linguistic Realization of Evidentiality in European Languages.* Berlin/New York: Mouton de Gruyter.

Cornillie, B., & Pietrandea, P. (2012). Modality at work. Cognitive, interactional and textual functions of modal markers. *Journal of Pragmatics*, *44*, 2109–2115.

Councill, I., McDonald, R., & Velikovich, L. (2010). What's Great and What's Not: Learning to Classify the Scope of Negation for Improved Sentiment Analysis. In *Proceedings of the NeSp-NLP Workshop* (pp. 51–59). Uppsala, Sweden.

Dahl, Ö. (1985). *Tense and Aspect Systems.* Oxford/New York: Blackwell Publishers.

De Haan, F. (1999). Evidentiality and epistemic modality: Setting boundaries. *Southwest Journal of Linguistics*, *18*, 83–101.

De Haan, F. (2006). Typological approaches to modality. In W. Frawley (Ed.), *The Expression of Modality.* Berlin/New York: Mouton de Gruyter.

Dowty, D. (1994). The Role of Negative Polarity and Concord Marking in Natural Language Reasoning. In *Proceedings of Semantics and Linguistics Theory (SALT)* (Vol. 4, pp. 114–144). Ithaca, USA.

Fauconnier, G. (1975). Pragmatic scales and logical structure. *Linguistic Inquiry*, *6*(3), 353–375.

Fernández Leborans, M. J. (1995). Las construcciones con el verbo estar: Aspectos sintácticos y semánticos [Constructions with verb estar: Syntactic and semantic aspects]. *Verba*, *22*, 252–284.

Fillmore, C. J. (1972). The Case for Case. In E. Bach & R. T. Harms (Eds.), *Universals in Linguistic Theory* (pp. 0–88). New York: Holt, Rinehart and Winston.

Frellesvig, B. (2010). *A History of the Japanese Language*. Cambridge: Cambridge University Press.

Fukushima, N. (2013a). *El español y el japonés [spanish and japanese]*. Kobe: Kobe University of Foreign Studies.

Fukushima, N. (2013b). 日西モダリティ対照研究史 [history of spanish and japanese contrastive studies]. *Kobe University Journal*, *63*(3), 3–11.

Garrote, M., Kimura, C., Matsui, K., Moreno, A., & Takamori, E. (2015). C-ORAL-JAPÓN: Corpus of Spontaneous Spoken Japanese. *Corpus Linguistics and Linguistic Theory*, *11*(2), 373–392.

Giammatteo, M., & Marcovecchio, A. M. (2010). Perífrasis verbales: Una mirada desde los universales lingüísticos. *Sintagma*, *21*, 21–38.

Gilligan, G. (1987). *Topic Continuity in Discourse: A Quantitative Cross-Language Study* (Unpublished doctoral dissertation). University of Southern California.

Gilquin, G., & de Cock, S. (2011). Errors and disfluencies in spoken corpora: Setting the scene. *International Journal of Corpus Linguistics*, *16*(2), 141–172.

Givón, T. (1995). *Functionalism and Grammar*. Amsterdam: John Benjamins Publishing.

Gómez Manzano, P. (1991). *Perífrasis Verbales con Infinitivo (Valores y Usos en la Lengua Hablada) [Infinitive Periphrastic Constructions. Meaning and Usage in the Spoken Language]*. Madrid: Universidad Nacional de Educación a Distancia.

Gómez Torrego, L. (1999). Los verbos auxiliares. Las perífrasis verbales de infinitivo. In *Gramática Descriptiva de la Lengua Española* (pp. 3323–3390). Madrid: Espasa Libros.

Grande Alija, F. (2002). *Aproximación a las Modalidades Enunciativas*. León: Universidad de León. Secretariado de Publicaciones y Medios Audiovisuales.

Gvozdanovic, J. (1989). Defining markedness. In O. M. Tomic (Ed.), *Markedness in Synchrony and Diachrony* (pp. 47–67). Berlin/New York: Mouton de Gruyter.

Halliday, M. (1970 [2009]). Functional diversity in language as seen from a consideration of modality and mood in English. In J. J. Webster (Ed.), *Studies in English Language. M.A.K. Halliday.* London: Continuum. (Originally published in Halliday (1970) Functional diversity in language as seen from a consideration of modality and mood in English.)

Harada, T. (1999). Modariti-ron shōkō – modariti o meguru nihongo kenkyū no futatsu no dōkō [Thoughts on modality theory – two tendencies in Japanese modality research]. *Genko to Bunka*, *3*, 123/136.

Hawkins, J. (1986). *A Comparative Typology of English and German.* London: Croom Helm.

Hennemann, A. (2013). *A Context-sensitive and Functional Approach to Evidentiality in Spanish or Why Evidentiality needs a Superordinate Category.* Pieterlen and Bern: Peter Lang, International Academic Publishers.

Herrero, C. (2013a). An initial approach on medical term formation in Japanese through the usage of corpora. In *Proceedings of the 7th Corpus Linguistics Conference* (pp. 339–340). Lancaster, United Kingdom.

Herrero, C. (2013b). A statistical study of the usage of no-negation and not-negation in spoken academic English. *Procedia - Social and Behavioral Science*, *95*, 482–489.

Herrero, C. (2014). Deontic Modality Markers in a Spanish Spoken Corpus. In *International Conference on Evidentiality and Modality in European Languages (EMEL 14).* Madrid.

Herrero, C., Campillos Llanos, L., & Moreno, A. (2014). Collecting and POS-tagging a lexical resource of Japanese biomedical terms from a corpus. *Revista de la Sociedad del Procesamiento de Lenguaje Natural (SEPLN)*, *52*, 29–36.

Herrero, C., & Moreno, A. (2014). The presence of modal periphrastic constructions in the Spanish formal and informal spoken discourse. In *Corpus linguistics international conference (CLIC 2014).* Las Palmas de Gran Canaria.

Hintikka, J. (1962). *Knowledge and Belief.* New York: Cornell University Press.

Hintikka, J. (2002). Negation in logic and in natural language. *Linguistics and Philosophy*, *25*(5/6).

Hisamitsu, T., & Nitta, Y. (1996). Analysis of Japanese compund nouns by direct text scanning. *Proceedings 16th Conference on Computational Linguistics*, *1*, 550–555.

Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization*. Cambridge: Cambridge University Press.

Horie, K. (2014). Modariti no ruikeiron [A typology of modal expressions]. In S. Harumi (Ed.), *Modariti: Riron to Hōhō [Modality: Theory and Method]*. Tokyo: Hitsuji Shobo.

Horie, K., & Narrog, H. (2014). What typology reveals about modality in Japanese. A cross-linguistic perspective. In K. Tabata & T. Ono (Eds.), *Usage-Based Approaches to Japanese Grammar. Towards the Understanding of Human Language.* Amsterdam/Philadelphia: John Benjamins Publishing.

Horn, L. (1989). *A Natural History of Negation*. Chicago: University Of Chicago Press.

Huddleston, R. D., & Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing In Science & Engineering*, *9*(3), 90–95.

Ide, S. (1992). The concept of politeness. An empirical study of American English and Japanese. In R. Watts, S. Ide, & K. Ehlich (Eds.), *Politeness in Language: Studies in its History, Theory and Practice* (pp. 281–297). Berlin/New York: Mouton de Gruyter.

Imithani, N. (2009). An overview of Japanese modalitites and their degree of proposition. *Humaniora*, *21*(1), 56–62.

Ingram, J., Hand, C. J., & Maciejewski, G. (2016). Exploring the Measurement of Markedness and Its Relationship with Other Linguistic Variables. *PloS one*, *11*(6).

Iori, I. (2014). Notes on the subjunctive mood in modern Japanese. *Hitotsubashi Journal of Arts and Sciences*, *55*(1), 45–57.

Iwasaki, S. (2013). *Japanese. Revised Edition.* Amsterdam/Philadelphia: John Benjamins Publishing.

Jiménez Juliá, T. (1989). Modalidad, modo verbal y modus clausal en español. *Verba. Anuario galego de filoloxía*, *16*, 175–214.

Johnson, Y. (1999). Modariti riron no meikakuka o motomete [Towards a clarification of modality theory]. In Y. Alam/Sazaki (Ed.), *Gengogaku to nihongo kyōiku* (p. 145/160). Tōkyō: Kuroshio Suppan.

Johnson, Y. (2003). *Modality and the Japanese Language.* Ann Arbor: The University of Michigan.

Kaiser, S., Ichikawa, Y., Kobayashi, N., & Yamamoto, H. (2013). *Japanese: A Comprehensive Grammar.* London/New York: Routledge.

Kataoka, K. (2012). 否定関連現象から見た日本語とスペイン語 [Syntactic Analysis of Negation Phenomena in Japanese and Spanish]. 日本語・日本学研究, *2*, 113–129.

Kaufmann, S., Condoravdi, C., & Harizanov, V. (2006). Formal approaches to modality. In W. Frawley (Ed.), *The Expression of Modality.* Berlin/New York: Mouton de Gruyter.

Kaul de Marlangeon, S. B. (2002). Los adverbios en -mente del español de hoy y su función semántica de cuantificación [Adverbs ended in -mente from today's Spanish and their cuantification semantic function.]. *Lingüística Iberoamericana*, *16*.

Kawazoe, A., Saitō, M., Kataoka, K., Choi, Y., & Bekki, D. (2010). Gengo jōhō no kakujitsusei anotēshon no tame no yōsō hyōgen no bunrui [Classification of the expression of aspect for the annotation of certainty information in language]. *Kyūshū daigaku gengogaku ronshū*, *31*(8), 109–129.

Kovacci, O. (1999). El adverbio. In *Gramática Descriptiva de la Lengua Española* (pp. 705–787). Madrid: Espasa Libros.

Kratzer, A. (1981). The notional category of modality. In H. J. Eikmeyer & H. Rieser (Eds.), *Words, Worlds, and Contexts. New Approaches in Word Semantics* (pp. 38–74). Berlin/New York: Mouton de Gruyter.

Kripke, S. (1963). Semantical analysis of modal logic. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, *9*, 67–96.

Kripke, S. A. (1963). Semantical Analysis of Modal Logic I. Normal Propositional Calculi. *Zeitschrift fur mathematische Logik und Grundlagen der Mathematik*, *9*(56), 67–96.

Kubler, S., McDonald, R., & Nivre, J. (2009). *Dependency Parsing.* San Rafael: Morgan and Claypool Publishers.

Kudo, T., & Matsumoto, Y. (2002). Japanese Dependency Analysis Using Cascaded Chunking. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20* (pp. 1–7). Stroudsburg, USA.

Kurohashi, S., & Nagao, M. (1994). KN Parser : Japanese Dependency/Case Structure Analyzer. In *In Proceedings of the Workshop on Sharable Natural Language Resources* (pp. 48–55).

Kurohashi, S., & Nagao, M. (1998). Building a Japanese Parsed Corpus while Improving the Parsing System. In *Proceedings of the NLPRS* (pp. 719–724).

Ladusaw, W. (1979). *Polarity Sensitivity as Inherent Scope Relations.* New York/London: Garland Publishing.

Lana-Serrano, S., Sánchez-Cisneros, D., Martínez, P., Moreno, A., & Campillos-Llanos, L. (2012). An Approach for Detecting Modality and Negation in Texts by Using Rule-Based Techniques. In *CLEF (Online Working Notes/Labs/Workshop).* Italy.

Larm, L. (2006). *Modality in Japanese* (Unpublished doctoral dissertation). University of Oxford, Oxford.

Larrega, P. (2009). Towards a typology of modality in language. In R. Salkie, P. Busuttil, & J. van der Auwera (Eds.), . Berlin/New York: Mouton de Gruyter.

Leech, G. (1997). Introducing corpus annotation. In R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus Annotation* (pp. 1–18). London: Longman.

Leech, G., & Smith, N. (1999). The use of tagging. In H. van Halteren (Ed.), *Syntactic Wordclass Tagging.* Dordrecht: Springer Science + Business Media.

Lloberes, M., & Castellón, I. (2010, May). Spanish FreeLing Dependency Grammar. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010).* La Valletta, Malta.

lxml Project. (2016). *Lxml Library.*

Lyons, J. (1977). *Semantics, Volume 2.* Cambridge: Cambridge University Press.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., … Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, *48*(2), 345–371.

Makino, S., & Tsutsui, M. (1994). *A Dictionary of Basic Japanese Grammar.* Tokyo: The Japan Times.

Makino, S., & Tsutsui, M. (1995). *A Dictionary of Intermediate Japanese Grammar.* Tokyo: The Japan Times.

Makino, S., & Tsutsui, M. (2008). *A Dictionary of Advanced Japanese Grammar.* Tokyo: The Japan Times.

Martín, S. E. (2004). *A Reference Grammar of Japanese.* Honolulu: University of Hawaii Press.

Masuoka, T. (1987). *Meidai no Bunpō [The Grammar of the Proposition].* Tokyo: Kuroshio.

Masuoka, T. (1991). *Modariti no Bunpou [The Grammar of the Modality].* Tokyo: Kuroshio.

Masuoka, T., & Takubo, Y. (1992). *Kiso nippon kataribunpō [Basic Japanese Grammar]*. Tokyo: Kuroshio Suppan.

Matsumoto, Y. (1989). Politeness and conversational universals, observations from Japanese. *Multilingua*, *8*(2/3), 207–222.

Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K., & Asahara, M. (2002). *Japanese Morphological Analysis System ChaSen Version 2.2.9 Manual* (Technical report, NAIST ed.).

Matsumoto, Y., Kurohashi, S., Yamaji, O., Taeki, Y., & Nagao, M. (1997). Japanese morphological analyzing system: JUMAN. *Technical Report, Kyoto University and Nara Institute of Science and Technology*.

Matsuoka, Y. (1981). *Handbook of Modern Japanese Grammar*. Tokyo: Hokuseido Press.

Matsuyoshi, S., Sato, S., & Utsuro, T. (2007). Nihongo kinō hyōgen jisho no hensan [Compilation of a dictionary of Japanese function expressions]. *Jizen Gengo Shori*, *14*(5), 123–146.

Maynard, S. K. (1993). *Discourse Modality: Subjectivity, Emotion and Voice in the Japanese Language (Pragmatics & Beyond New Series)*. Amsterdam: John Benjamins Publishing Company.

McCarthy, M. (1998). *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

Mcdonald, R., Nivre, J., Quirmbach-brundage, Y., Goldberg, Y., Das, D., Ganchev, K., … Lee, C. J. (2013). Universal dependency annotation for multilingual parsing. In *In Proc. of ACL '13.*

McEnery, A., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

McEnery, A., Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies*. London/New York: Routledge.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 51 – 56). Austin, USA.

Miller, J., & Weinert, R. (1998). *Spontaneous Spoken Language. Syntax and Discourse.* Oxford: Clarendon Press.

Morante, R., & Daelemans, W. (2012). Annotating modality and negation for a machine reading evaluation. In *Workshop on question answering for machine reading evaluation (qa4mre).* Rome, Italy.

Moreno, A., de la Madrid, G., Alcántara, M., González, A., & de la Torre, R. (2005). The Spanish corpus. In Cresti & Mongelia (Eds.), *C-ORAL-ROM Integrated Reference Corpora for Spoken Romance Languages.* Amsterdam: John Benjamins Publishing.

Moreno, A., & Goñi, J. M. (1995). A morphological model and processor for Spanish implemented in Prolog. In *Proceedings of Joint Conference on Declarative Programming* (pp. 321–331). Salerno, Italy.

Moreno, A., & Guirao, J. M. (2003). Tagging a spontaneous speech corpus of Spanish. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (pp. 292–296). Borovets, Bulgaria.

Moreno, A., & Guirao, J. M. (2006). Morpho-syntactic Tagging of the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation. In Y. Kawaguchi, S. Zaima, & T. Takagaki (Eds.), *Spoken Language Corpus and Linguistic Informatics.* Amsterdam: John Benjamins Publishing.

Moreno Cabrera, J. C. (2000). *Curso Univeristario de Linguistica General (Tomo 1: Teoria de la Gramatica y Sintaxis General) (Spanish Edition).* Madrid: Sintesis Editorial.

Moriya, T., & Horie, K. (2009). What Is and Is not Language-Specific about the Japanese Modal System? A Comparative and Historical Perspective. In B. Pizziconi & M. Kizu (Eds.), *Japanese Modality: Exploring its Scope and Interpretation.* London: Palgrave Macmillan.

Murata, M., Uchimoto, K., Ma, Q., & Isahara, H. (2001). Magical Number Seven Plus or Minus Two: Syntactic Structure Recognition in Japanese and English Sentences. In *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing* (pp. 43–52). London, United Kingdom.

Murawaki, Y., & Kurohashi, S. (2008). Online aquisition of Japanese unknown morphemes using morphological constraints. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 429–437). Honolulu, USA.

Nariyama, S. (2003). *Ellipsis and Reference Tracking in Japanese.* Amsterdam/Philadelphia: John Benjamins.

Narrog, H. (2009a). *Modality in Japanese. The Layered Structure of the Clause and Hierarchies of Functional Categories.* Amsterdam: John Benjamins Publishing.

Narrog, H. (2009b). Modality, Modariti and Predication: The Story of Modality in Japan. In B. Pizziconi & M. Kizu (Eds.), *Japanese Modality: Exploring its Scope and Interpretation.* London: Palgrave Macmillan.

Narrog, H. (2012). *Modality, Subjectivity and Semantic Change.* Oxford: Oxford University Press.

Nitta, Y. (1985). Bun no honegumi [The frame of a sentence]. In H. Shirō (Ed.), *Nihongo no Kyōiku.* Tokyo: Meijishoin.

Nitta, Y. (1991). *Nihongo no Modariti to Ninshou.* Tokyo: Hitsuji Shobo.

Nivre, J. (2015). Towards a Universal Grammar for Natural Language Processing. In *Proceedings of the 16th international conference of computational Linguistics and Intelligent Text Processing (cicling 2015)* (pp. 3–16). Cairo, Egipt.

Nomura, K. (2010). *Japanese Grammar: The Connecting Point.* London/Plymouth: University Press of America.

Nomura, T. (2003). Modariti keisiki no bunrui [On the classification of Japanese modal forms]. *Kokugogaku, 54 (1)*, 17–31.

Nuyts, J. (2001). *Epistemic Modality, Language, and Conceptualization.* Amsterdam: John Benjamins Publishing.

Nuyts, J. (2006). Modality: Overview and linguistic issues. In W. Frawley (Ed.), *The Expression of Modality. The Expression of Cognitive Categories.* Berlin/New York: Mouton de Gruyter.

Obana, Y. (1997). Vertical or Horizontal? Reading Directions in Japanese. *Bulletin of the School of Oriental and African Studies, University of London*, *60*(1), 86–94.

O'Connell, C., Daniel, & Kowal, S. (2005, November). Uh and Um Revisited: Are They Interjections for Signaling Delay? *Journal of psycholinguistic research*, *34*(6), 555–76. (Copyright - Springer Science+Business Media, Inc. 2005; Última actualización - 2014-07-27)

Onoe, K. (1990). Bunpōron: Chinjutsu-ron no tanjō to shūen [Grammar: The birth and the death of predication theory]. *Kokugo to Kokubungaku*, *67*(4), 1–16.

Onoe, K. (2004). Shugo to jutsugo wo meguru bunpō [Grammar concerning subject and predicate]. In K. Onoe (Ed.), *Asakura Nihongo Kōza 6: Bunpō II [Asakura Series on Japanese Language Volume 6: Grammar II].* Tokyo: Asakura Syoten.

Osborne, T. (2014). Dependency Grammar. In A. Carnie, Y. Sato, & D. Siddiqi (Eds.), *The Routledge Handbook of Syntax.* London/New York: Routledge.

Otaola Olano, C. (1988). La modalidad (con especial referencia a la lengua española). *Revista de Filología Española*, *68*(1), 97–117.

Pakray, P., Bhaskar, P., Banerjee, S., Bandyopadhyay, S., & Gelbukh, A. (2012). An automatic system for modality and negation detection. In *Workshop on question answering for machine reading evaluation (qa4mre).* Rome, Italy.

Palmer, F. (2001). *Mood and modality.* Cambridge: Cambridge University Press.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, *31*(1), 71–106.

Patterson, R. (2002). *Aristotle's Modal Logic: Essence and Entailment in the Organon.*

Piao, S., Bianchi, F., Dayrell, C., D'Egidio, A., & Rayson, P. (2015). Development of the multilingual semantic annotation system. In *Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics - human language technologies (naacl hlt 2015).* Denver, USA.

Python Software Foundation. (2016). *Python Language reference, version 3.6.*

Quirk, R., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language.* London: Longman.

Radden, G. (2014). Making sense of negated modals. *Argumentum*, *10*, 519–532.

RAE. (2001). *Diccionario de la Lengua Española [Dictionary of Spanish Language].* Madrid: Espasa Libros.

RAE. (2009). *Nueva Gramática de la Lengua Española [New Grammar of the Spanish Language].* Madrid: Espasa Libros.

Rayson, P., Archer, D., Piao, S., & McEnery, T. (2004). The UCREL semantic analysis system. In *Proceedings of the workshop on beyond named entity recognition semantic labelling for nlp tasks in association with 4th international conference on language resources and evaluation (LREC 2004).* Lisbon, Portugal.

Richardson, L. (2016). *Beautiful Soup.*

Ringbom, H. (1987). *The Role of the First Language in Foreign Language Learning.* Philadelphia: Multilingual Matters.

Rini, A. (2011). *Aristotle's Modal Proofs: Prior Analytics A8-22 in Predicate Logic* (1st ed.). Amsterdam: Springer.

Rodríguez Ramalle, T. M. (2003). *La Gramática de los Adverbios en -mente o Cómo Expresar Maneras, Opiniones y Actitudes a Través de la Lengua [The Grammar of the Adverbs Ended in -mente, or How to Express Manner, Opinions and Attitudes Through Language.].* Madrid: UAM Ediciones.

Rosenberg, S., Kilicoglu, H., & Bergler, S. (2012). Clac labs processing modality and negation working notes for qa4mre pilot task at clef 2012. In *Workshop on question answering for machine reading evaluation (qa4mre).* Rome, Italy.

Sanz Alonso, B. (1996). *La Negación en Español*. Salamanca: Ediciones Colegio de España.

Saurí, R., & Pustejovsky, J. (2009). FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation*, *43*(3), 227–268.

Scarcela, R., & Brunak, J. (1981). On speaking politely in a second language. *International Journal of the Sociology of Language*, 59–75.

Shibatani, M. (1990). *The Languages of Japan*. Cambridge: Cambridge University Press.

Squartini, M. (2004). Disentangling evidentiality and epistemic modality in romance. *Lingua*, *114*(7), 873–889.

Srdanovic, I., Suchomel, V., Ogiso, T., & Kilgarriff, A. (2013). Hyakuokugo no kōpasu wo mochiita nihongo no goi, bunpō jōhō no purofairingu [Japanese Language Lexical and Grammatical Profiling Using the Web Corpus JpTenTen]. In *Daisankai kōpasu nihongaku wōkushoppu yokōsū [Proceedings of the 3rd Japanese corpus linguistics workshop]* (pp. 229–238). Tokyo, Japan.

Stowell, T. (2004). Tense and modals. In J. Gueron & J. Lecarme (Eds.), *The Syntax of Time* (pp. 621–636). Cambridge MA: MIT Press.

Stubbs, M. (1986). A matter of prolonged field work: Notes towards a modal grammar of English. *Applied Linguistics*, *7*, 1–25.

Swan, T. (1991). Adverbial shifts: Evidence from Norwegian and English. In *Historical English Syntax.* Berlin/New York: Mouton de Gruyter.

Talué, M., Martí, A., & Recasens, M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008).* Marrakech, Morocco.

Tanaka, T., Miyao, Y., Asahara, M., Uematsu, S., Kanayama, H., Mori, S., & Matsumoto, Y. (2016). Universal Dependencies for Japanese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 23–28). Portorož, Slovenia.

Tenfjord, K., Hagen, J., & Johansen, H. (2006). The hows and whys of coding categories in a learner corpus (or 'how and why an error-tagged learner corpus is not ipso facto one big comparative fallacy'). *Rivista di Psicolinguistica Applicata (RiPLA)*, *6*(3), 93–108.

Tomanek, K., Wermter, J., & Hahn, U. (2007). A reappraisal of sentence and token splitting for life sciences documents. *Studies In Health Technology And Informatics*, *129*(Pt 1), 524 – 528.

Traugott, E. C. (2006). Historical aspects of modality. In W. Frawley (Ed.), *The Expression of Modality* (pp. 107–141). Berlin/New York: Mouton de Gruyter.

Uchiyama, K., Baldwin, T., & Ishizaki, S. (2005). Disambiguating Japanese compund verbs. *Computer Speech and Language*, *19*(4), 497–512.

Van der Auwera, J. (1985). *Language and Logic. A Speculative and Condition-Theoretic Study.* Amsterdam/Philadelphia: John Benjamins Publishing.

Van der Auwera, J., & Ammann, A. (2013). Overlap between Situational and Epistemic Modal Marking. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Van der Auwera, J., Kehayov, P., & Vittrant, A. (2009). Acquisitive modals. In *Crosslinguistic Semantics of Tense, Aspect and Modality.* Amsterdam: Benjamins.

Van der Auwera, J., & Plungian, V. A. (1998). Modality's semantic map. *Linguistic Typology*, *2*, 79–124.

Van der Auwera, J., & Zamorano Aquilar, A. (2015). The History of Modality and Mood. In J. Nuyts & J. Van der Auwera (Eds.), *The Oxford Handbook of Modality and Mood.* Oxford: Oxford University Press.

van Valin, R. D., & LaPolla, R. J. (1997). *Syntax: Structure, meaning and function.* Cambridge: Cambridge University Press.

Von Fintel, K. (2006). Modality and Language. In D. M. Borchert (Ed.), *Encyclopedia of Philosophy.* Detroit: MacMillan Reference USA.

Von Wright, E. (1951). *An Essay in Modal Logic.* Amsterdam: North Holland.

Voutilainen, G. (1999). Orientation. In H. van Halteren (Ed.), *Syntactic Wordclass Tagging.* Dordrecht: Springer Science + Business Media.

Wasa, A. (2005). *Supeingo to nihongo no modariti [Japanese and Spanish modality].* Tokyo: Kuroshio.

Wilson, T., Wiebe, J., & Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, *35*(3), 399–433.

Zeman, S. (2014). (C)Overt epistemic modality and its perspectival effects on the textual surface. In *Modes of Modality: Modality, Typology and Universal Grammar.* Amsterdam: John Benjamins Publishing Company.

Appendix A

Frequencies

Table 104: Spanish necessity, possibility, epistemic, deontic and ambiguous markers absolute frequencies

Speaker	NEC	POSS	EPIS	DEON	AMBG	Words
ABE_man_Madrid	17	5	1	14	7	1437
ABE_man_C_x	9	0	1	8	0	663
ABU_woman_D_Madrid	19	10	3	10	16	2773
ADI_man_Morocco	10	14	4	8	12	1071
ADR_woman_B_Colombia	1	2	0	0	3	727
AFR_woman_x_x	0	0	0	0	0	1
AIT_woman_Madrid	0	0	0	0	0	2
ALB_man_B_Segovia	1	1	0	1	1	313
ALB_man_C_Madrid	15	8	1	8	14	873

ALF_man_C_Segovia	6	2	2	5	1	1046
ALI_woman_B_Mexico	1	1	1	1	0	175
ALI_woman_C_x	5	2	0	3	4	544
ALM_woman_B_Madrid	3	4	2	3	2	411
ALO_man_x_x	1	0	0	1	0	164
ALV_man_Madrid	32	13	3	29	13	2233
ALV_woman_x_x	2	1	1	1	1	473
AMA_woman_Madrid	11	10	10	9	2	834
ANA_woman_Madrid	14	8	1	15	6	957
ANA_woman_B_Madrid	14	8	3	8	11	1470
ANA_woman_B_Salamanca	18	1	0	17	2	858
ANA_woman_C_x	10	8	1	9	8	831
ANA_woman_x_x	2	0	0	2	0	126
AND_man_C_x	1	1	0	1	1	251
ANG_man_B_Madrid	26	10	5	21	10	3019
ANG_man_C_Segovia	32	11	9	26	8	3001

ANG_man_C_x	12	10	14	4	4	1630
ANS_man_C_x	2	0	0	1	1	511
ANT_man_B_Madrid	28	10	8	19	11	2300
ANT_man_x_x	0	0	0	0	0	80
ANU_man_x_x	0	0	0	0	0	7
APU_man_B_Segovia	0	0	0	0	0	13
ARA_woman_Madrid	1	2	1	2	0	759
ARA_woman_B_Madrid	10	5	1	8	6	863
ARA_woman_C_x	4	2	0	4	2	273
ARI_man_xñrgentina	8	6	1	5	8	616
ART_man_x_x	0	0	0	0	0	29
ATR_woman_x_x	0	0	0	0	0	37
AVE_man_x_x	7	6	3	0	10	1282
BAJ_man_x_x	0	0	0	0	0	65
BAR_woman_Madrid	2	4	3	0	3	411
BEA_woman_Navarra	1	0	0	1	0	116

BEA_woman_B_Navarra	28	12	3	16	21	2516
BEC_woman_Madrid	5	5	2	3	5	846
BEL_man_B_Madrid	11	5	7	6	3	863
BEL_woman_B_Madrid	5	4	0	5	4	830
BEL_woman_x_x	0	0	0	0	0	23
BIE_man_x_x	0	0	0	0	0	21
BLA_man_x_x	4	4	0	4	4	581
BLA_woman_Madrid	3	1	0	3	1	138
BRE_man_x_x	1	1	0	0	2	214
BUS_man_C_x	0	1	0	0	1	265
CAL_man_x_x	0	0	0	0	0	6
CAM_man_C_Murcia	1	0	0	1	0	71
CAM_woman_B_x	0	0	0	0	0	113
CAR_man_Segovia	6	0	0	6	0	89
CAR_man_xñlicante	0	0	0	0	0	54
CAR_man_x_Madrid	1	0	0	1	0	345

CAR_woman_B_Madrid	18	18	3	11	22	1064
CAR_woman_C_Madrid	31	19	6	16	28	2568
CAR_woman_C_x	1	0	0	1	0	194
CAY_man_x_x	2	0	0	2	0	75
CEL_woman_x_x	0	1	0	0	1	226
CES_man_B_Segovia	3	2	1	3	1	461
CES_man_x_x	0	7	4	1	2	697
CHA_man_B_Valladolid	29	5	2	12	20	2990
CHA_woman_x_x	1	1	0	1	1	64
CHE_man_Ferrol	0	0	0	0	0	0
CHI_man_B_Madrid	0	0	0	0	0	435
CHI_woman_x_x	1	0	0	1	0	122
CIE_man_C_x	7	11	8	3	7	1071
CLE_woman_C_Segovia	4	0	0	4	0	299
COB_man_x_x	0	1	1	0	0	139
CON_woman_x_x	0	0	0	0	0	86

CRI_man_B_Madrid	19	12	7	9	15	3011
CRI_woman_Madrid	6	1	0	3	4	757
CRI_woman_Segovia	6	1	0	4	3	731
CRI_woman_B_Madrid	11	18	11	7	11	2256
CRI_woman_B_Segovia	0	0	0	0	0	13
CUR_man_x_x	1	0	0	1	0	246
DAM_man_x_x	1	4	0	1	4	163
DAN_man_Bñrgentina	3	0	2	1	0	725
DAN_man_B_Madrid	5	0	0	5	0	305
DAN_man_x_x	0	0	0	0	0	172
DAS_woman_B_Madrid	20	0	0	17	3	1018
DAV_man_Madrid	40	9	8	30	11	4581
DAV_man_B_Segovia	4	1	0	4	1	232
DAV_man_x_x	0	0	0	0	0	46
DEB_man_B_Madrid	4	5	2	2	5	486
DEF_man_C_Cataluña	24	4	2	18	8	1273

DIR_man_x_x	0	0	0	0	0	63
DOA_woman_Egipto	2	1	1	2	0	219
DOC_man_x_x	0	0	0	0	0	9
DOL_woman_D_Madrid	20	1	1	20	0	1116
DOS_man_C_Madrid	26	16	5	13	24	1822
DOS_man_x_Mexico	2	2	0	3	1	132
DRA_man_x_x	1	2	2	0	1	133
DUD_man_B_Segovia	0	0	0	0	0	0
DUR_man_C_Cataluña	0	3	1	0	2	98
EDU_man_C_x	6	7	2	3	8	846
ELA_woman_x_x	2	0	0	2	0	189
ELE_woman_Madrid	16	6	2	15	5	1327
ELE_woman_B_Madrid	5	2	3	3	1	1048
ELE_woman_C_x	0	0	0	0	0	115
ELI_man_x_x	1	0	1	0	0	55
EMB_man_x_x	3	3	3	1	2	491

EMI_man_Madrid	3	7	4	2	4	688
EMI_man_x_x	1	2	0	1	2	243
ENF_woman_x_x	0	0	0	0	0	29
ENR_man_B_Madrid	17	15	5	11	16	3834
ENT_man_x	3	1	0	1	3	177
ESQ_man_x_x	0	0	0	0	0	55
EST_man_B_x	0	0	0	0	0	58
EST_man_C_x	3	2	1	1	3	312
EST_man_xñndalucñ	5	1	1	4	1	365
EST_woman_B_Madrid	7	0	0	7	0	568
EST_woman_x_x	0	1	0	0	1	106
EVA_woman_Madrid	7	0	1	5	1	608
EVO_man_B_Madrid	4	1	0	2	3	515
FEL_man_x_x	0	0	0	0	0	2
FER_man_Madrid	3	0	0	2	1	391
FER_man_x_x	15	2	2	11	4	1395

FIS_man_C_x	12	3	1	12	2	985
FRA_man_x_x	2	1	0	0	3	114
FRE_man_x_x	0	0	0	0	0	12
FUE_man_B_Madrid	7	3	2	3	5	1031
FUE_woman_Madrid	3	3	3	1	2	371
FUE_woman_C_x	0	0	0	0	0	0
FUE_woman_x_x	4	5	2	2	5	1201
GAB_man_x_x	11	1	1	4	7	622
GAN_man_C_x	0	0	0	0	0	96
GAR_man_x_x	8	3	1	6	4	2041
GAS_man_C_x	0	0	0	0	0	63
GAT_man_x_x	0	0	0	0	0	91
GEM_woman_Madrid	13	0	0	10	3	701
GEM_woman_B_Madrid	19	5	5	10	9	690
GLO_woman_B_Madrid	4	6	2	4	4	959
GRA_man_x_x	1	2	0	1	2	117

GUA_man_B_Barcelona	3	0	0	3	0	108
GUI_man_Madrid	9	4	2	6	5	1140
GUI_man_B_Madrid	10	0	2	5	3	520
GUI_man_x_x	3	0	0	2	1	120
GUS_manñrgentina	3	0	0	2	1	824
GUT_man_x_x	5	1	0	6	0	213
HAB_woman_x_x	3	3	0	2	4	111
HEC_man_x_x	0	0	0	0	0	125
HEI_man_C_Madrid	1	0	0	1	0	376
HEL_woman_Madrid	46	7	0	42	11	3584
HER_man_B_Madrid	30	19	5	26	18	2924
HER_man_x_x	0	0	0	0	0	0
HER_woman_x_x	0	0	0	0	0	2
HIS_man_Madrid	12	9	7	7	7	4509
HOM_man_C_Madrid	7	2	1	6	2	805
HOM_man_x_x	1	0	0	1	0	177

HOY_man_B_Burgos	13	15	7	9	12	4489
IDO_woman_B_Madrid	7	4	0	7	4	319
INM_woman_Madrid	15	7	3	9	10	1895
INM_woman_x_x	1	0	0	1	0	46
ISA_woman_C_Madrid	1	1	0	1	1	254
ISA_woman_x_x	0	1	0	0	1	137
ISM_man_C_x	0	0	0	0	0	120
IVA_man_Cantabria	0	0	0	0	0	76
JAI_man_C_x	1	0	0	0	1	60
JAU_man_x_x	0	3	3	0	0	115
JAV_man_B_Madrid	0	6	3	0	3	356
JAV_man_B_Palencia	14	12	8	12	6	1133
JAV_man_x_x	2	1	1	1	1	331
JES_man_Madrid	0	0	0	0	0	1
JES_man_B_Segovia	21	5	3	9	14	2224
JES_man_C_x	2	2	0	0	4	475

JOA_man_Madrid	65	21	9	49	28	3883
JOA_man_x_Cataluña	0	0	0	0	0	197
JOA_man_x_Extremadura	3	1	3	0	1	333
JOS_man_Segovia	4	1	1	1	3	378
JOS_man_B_Madrid	31	35	11	21	34	4410
JOS_man_C_Madrid	44	46	18	28	44	8304
JOS_man_C_Segovia	19	10	5	16	8	3034
JOS_man_D_Madrid	1	0	0	1	0	102
JOV_woman_B_Madrid	7	1	1	6	1	722
JOV_woman_x_Mexico	1	0	0	1	0	60
JUA_man_C_Salamanca	3	0	0	3	0	216
JUA_man_C_x	16	12	6	14	8	1336
JUA_woman_x_x	0	0	0	0	0	22
JUE_man_C_x	5	1	0	5	1	200
JUL_man_C_x	16	22	6	5	27	1581
KAJ_man_x_x	0	1	0	0	1	104

LAG_man_C_Madrid	1	2	0	1	2	602
LAN_woman_Madrid	15	4	3	11	5	829
LAR_man_x_x	0	0	0	0	0	6
LAU_woman_B_Madrid	7	4	1	6	4	592
LAU_woman_x_x	2	0	0	2	0	323
LET_woman_B_Madrid	17	7	0	13	11	1247
LIN_woman_x_x	1	0	0	1	0	74
LIS_man_x_x	0	7	4	0	3	533
LOC_man_C_Madrid	44	26	8	32	30	3879
LOC_man_C_x	15	2	2	12	3	1983
LOD_man_x_x	0	0	0	0	0	3
LOL_woman_B_Madrid	3	1	0	2	2	872
LOL_woman_C_x	0	0	0	0	0	334
LOL_woman_x_x	0	1	1	0	0	43
LUC_man_B_Zamora	24	13	3	23	11	1521
LUC_man_C_x	1	1	0	1	1	134

LUC_woman_Madrid	17	10	3	13	11	1346
LUI_man_Madrid	10	12	5	9	8	683
LUI_man_C_Madrid	10	3	2	10	1	723
LUI_man_C_Segovia	4	1	1	2	2	470
LUI_man_C_x	5	8	3	4	6	561
LUI_woman_B_Burgos	8	12	5	7	8	3115
LUI_woman_C_Burgos	5	3	1	5	2	503
LUI_woman_D_Madrid	1	0	0	1	0	25
LUZ_woman_C_Madrid	3	0	0	3	0	268
MAC_man_x_Mexico	0	0	0	0	0	64
MAD_man_x_Mexico	0	0	0	0	0	95
MAD_woman_C_Badajoz	17	1	0	13	5	1051
MAD_woman_C_Madrid	1	0	0	1	0	44
MAD_woman_C_Segovia	11	1	1	9	2	1041
MAF_woman_Madrid	26	13	3	16	20	3190
MAJ_man_x_Mexico	1	0	0	1	0	154

MAL_man_x_x	0	1	0	0	1	47
MAM_woman_C_Madrid	36	1	1	26	10	1963
MAM_woman_C_Segovia	11	0	0	11	0	128
MAN_man_Madrid	10	6	0	6	10	1177
MAN_man_C_Madrid	19	5	3	12	9	1229
MAN_man_x_x	5	4	3	4	2	1170
MAR_man_B_Madrid	9	3	1	8	3	463
MAR_man_C_Madrid	23	12	9	12	14	3040
MAR_man_C_x	15	10	2	14	9	1211
MAR_woman_Madrid	60	7	5	48	14	6286
MAR_woman_B_Madrid	6	1	0	4	3	315
MAR_woman_B_x	0	0	0	0	0	21
MAR_woman_C_Madrid	14	4	3	3	12	2293
MAR_woman_D_Madrid	41	4	2	30	13	4579
MAR_woman_x_x	3	5	2	1	5	1363
MAS_man_x_x	0	0	0	0	0	12

MAS_woman_Madrid	0	0	0	0	0	24
MAU_man_D_Burgos	16	5	3	14	4	1692
MAY_man_C_x	0	0	0	0	0	61
MAY_woman_B_Madrid	21	3	2	9	13	844
MAY_woman_C_Badajoz	35	7	4	24	14	4490
MAY_woman_x_x	0	0	0	0	0	39
MER_man_x_x	0	0	0	0	0	49
MIG_man_Madrid	11	1	1	11	0	795
MIG_man_B_Madrid	28	11	7	22	10	3335
MIG_man_C_Segovia	0	0	0	0	0	0
MIG_man_x_x	0	0	0	0	0	547
MIL_man_x_x	0	0	0	0	0	30
MOD_man_C_x	0	0	0	0	0	74
MOD_man_x_x	3	0	0	3	0	117
MON_man_C_Madrid	20	9	8	7	14	1644
MON_woman_Madrid	24	6	2	19	9	2567

MON_woman_B_Madrid	42	8	4	33	13	4550
MOR_man_C_ceres	0	1	0	0	1	226
MUJ_woman_B_Madrid	0	0	0	0	0	0
MUJ_woman_x_x	0	1	0	0	1	259
NAD_man_B_Mallorca	0	0	0	0	0	34
NAN_woman_Madrid	7	1	1	5	2	720
NAR_man_x_x	0	1	0	0	1	703
NAT_woman_B_Madrid	6	4	1	6	3	572
NEN_woman_Madrid	7	0	0	6	1	737
NIE_woman_B_Madrid	3	0	0	2	1	178
NIE_woman_C_Madrid	6	7	3	2	8	736
NIV_woman_C_Madrid	9	0	0	5	4	491
NOT_man_x_x	4	1	2	2	1	532
NUR_woman_B_Madrid	15	7	4	12	6	1675
OLG_woman_Madrid	1	0	1	0	0	168
OLG_woman_B_Madrid	5	9	5	4	5	980

OTR_man_x_x	0	1	0	0	1	298
OTR_woman_Madrid	2	0	0	2	0	98
OTR_woman_x_x	5	2	0	2	5	642
OSO_man_B_Madrid	3	1	2	1	1	724
PAC_man_Madrid	7	11	9	5	4	1824
PAC_man_B_Sevilla	11	1	0	6	6	1368
PAC_man_x_x	6	3	1	5	3	741
PAD_man_C_Madrid	2	0	1	1	0	40
PAL_woman_Madrid	5	0	0	4	1	362
PAL_woman_B_Madrid	8	0	1	5	2	431
PAN_man_x_x	5	9	5	1	8	1265
PAP_man_C_Madrid	1	1	0	1	1	232
PAP_man_C_Segovia	0	0	0	0	0	17
PAR_man_x_x	0	0	0	0	0	3
PAS_man_B_Madrid	3	2	3	0	2	248
PAS_woman_x_x	0	0	0	0	0	77

PAT_man_B_x	0	0	0	0	0	3
PAT_woman_Madrid	15	8	6	9	8	1539
PAT_woman_B_Madrid	20	8	7	15	6	1380
PAT_woman_B_x	0	0	0	0	0	5
PAT_woman_C_x	0	0	0	0	0	191
PAZ_woman_Madrid	9	1	1	7	2	1371
PED_man_x_x	0	6	1	0	5	785
PEN_man_C_x	1	0	0	1	0	116
PEP_man_x_x	3	0	0	1	2	173
PEP_woman_C_Segovia	9	0	0	0	9	1206
PEP_woman_x_x	6	1	0	6	1	535
PEQ_man_x_x	0	0	0	0	0	141
PES_woman_C_Madrid	5	0	0	5	0	525
PFE_man_x_x	0	0	0	0	0	25
PIB_man_x_x	0	1	0	0	1	359
PIE_woman_D_Madrid	6	1	2	2	3	803

PIL_woman_B_Ciudad Real	7	3	3	6	1	935
PIL_woman_B_Segovia	10	2	1	7	4	1647
PIL_woman_C_Madrid	4	5	1	1	7	632
PIL_woman_x_x	4	2	2	2	2	1076
PIM_man_x_x	0	0	0	0	0	27
PLA_man_C_x	2	3	0	2	3	245
PLA_woman_x_x	1	1	0	0	2	428
POL_man_C_País Vasco	7	3	3	2	5	1133
PRE_man_C_x	21	13	7	14	13	2450
PRI_man_x_Mexico	0	1	0	0	1	99
PRI_woman_B_Madrid	1	0	0	1	0	309
PRI_woman_C_Madrid	64	14	6	47	25	5501
PRI_woman_x_x	0	0	0	0	0	6
PRO_man_x_x	2	0	0	2	0	62
QUI_man_Madrid	6	4	2	5	3	794
QUI_man_x_x	2	1	0	1	2	85

RAF_man_C_x	11	9	5	6	9	1361
RAM_man_x_x	1	1	0	1	1	238
RAT_man_C_x	1	0	0	1	0	115
RAU_man_Madrid	32	2	0	25	9	1890
RAU_man_B_Madrid	2	2	0	2	2	710
RED_man_C_x	0	0	0	0	0	0
REP_woman_x_x	1	0	0	1	0	108
RES_man_x_x	1	0	0	1	0	322
RIC_man_B_Madrid	2	2	0	2	2	413
RIC_man_B_Segovia	27	5	3	21	8	3082
ROB_man_B_Segovia	18	3	2	15	4	1003
ROD_man_C_x	0	3	1	0	2	182
ROS_woman_B_Madrid	26	8	9	15	10	2078
SAL_woman_Madrid	2	1	1	2	0	192
SAN_woman_B_Madrid	0	0	0	0	0	0
SAR_man_x_x	0	0	0	0	0	27

SAR_woman_Madrid	10	5	4	6	5	633
SEC_man_B_Madrid	11	4	3	12	0	982
SEC_man_C_x	4	0	3	0	1	645
SEC_man_x_x	1	0	0	1	0	23
SER_man_B_Huelva	3	0	0	2	1	323
SER_man_B_Segovia	50	38	9	19	60	3476
SER_man_C_Barcelona	1	0	0	1	0	60
SER_man_x_Mexico	0	1	1	0	0	131
SEV_man_C_Valencia	4	2	2	3	1	974
SEV_man_C_Valladolid	1	0	0	1	0	159
SEV_woman_Sevilla	8	3	2	4	5	1157
SEV_woman_B_Sevilla	1	0	0	1	0	50
SEW_woman_x_x	0	0	0	0	0	116
SIG_man_D_Barcelona	9	28	14	3	20	3129
SIL_woman_B_Madrid	7	9	5	6	5	531
SOF_woman_Bñsturias	4	1	1	4	0	559

SON_woman_B_Madrid	15	7	6	14	2	795
SUS_woman_B_Valladolid	18	21	8	9	22	2995
TEL_man_x_x	2	0	0	2	0	54
TER_woman_B_Madrid	1	0	0	1	0	82
TIA_woman_B_Madrid	15	0	1	13	1	932
TIA_woman_x_x	0	0	0	0	0	5
TIE_man_x_x	0	0	0	0	0	518
TIO_man_x_x	0	0	0	0	0	1
TOD_woman_B_Madrid	0	0	0	0	0	1
TOM_man_x_x	0	1	0	0	1	249
TOR_man_C_x	2	6	3	2	3	656
TRA_man_x_x	0	1	0	0	1	157
TRE_man_x_Mexico	0	0	0	0	0	73
TRI_woman_Madrid	8	3	3	5	3	1032
UEL_man_B_Madrid	6	2	1	5	2	805
UNO_man_C_Madrid	4	2	0	4	2	1224

USE_man_B_Madrid	1	1	1	0	1	123
VAM_man_B_Madrid	0	0	0	0	0	0
VAY_man_x_Mexico	0	1	1	0	0	73
VER_woman_Madrid	7	5	1	5	6	1780
VIC_man_x	2	0	0	0	2	176
VIC_man_C_x	2	2	2	1	1	710
VIC_woman_C_Madrid	13	5	3	9	6	661
VID_man_C_Segovia	10	0	0	10	0	235
VIS_woman_C_Cuenca	19	0	0	15	4	1510
VIT_woman_B_Madrid	4	1	0	3	2	538
VRI_man_x_x	3	3	4	1	1	684
WAL_man_x_x	2	0	1	1	0	129
WOM_woman_x_x	0	0	0	0	0	11
XYZ_man_x_x	0	0	0	0	0	4
YOL_woman_Madrid	2	1	0	1	2	182
YOL_woman_B_Segovia	0	0	0	0	0	104

YUS_man_B_Segovia	1	0	0	1	0	265
ZAS_man_C_Madrid	4	0	0	4	0	1186
Total	2698	1253	594	1909	1448	301329
Mean	7.12	3.31	1.57	5.03	3.82	795.06

Table 105: Spanish markers' grammatical class absolute frequencies

Speaker	**AUX**	**ADV**	**ADJ**	**MOOD**
ABE_man_A_Madrid	16	1	0	5
ABE_man_C_x	7	1	0	1
ABU_woman_D_Madrid	22	1	1	5
ADI_man_A_Morocco	20	3	1	0
ADR_woman_B_Colombia	3	0	0	0
AFR_woman_x_x	0	0	0	0
AIT_woman_A_Madrid	0	0	0	0
ALB_man_B_Segovia	1	0	0	1
ALB_man_C_Madrid	21	1	0	1

ALF_man_C_Segovia	7	1	0	0
ALI_woman_B_Mexico	0	1	0	1
ALI_woman_C_x	6	0	1	0
ALM_woman_B_Madrid	6	1	0	0
ALO_man_x_x	1	0	0	0
ALV_man_A_Madrid	34	3	0	8
ALV_woman_x_x	2	1	0	0
AMA_woman_A_Madrid	13	4	2	2
ANA_woman_A_Madrid	20	1	0	1
ANA_woman_B_Madrid	17	3	0	2
ANA_woman_B_Salamanca	19	0	0	0
ANA_woman_C_x	17	1	0	0
ANA_woman_x_x	2	0	0	0
AND_man_C_x	2	0	0	0
ANG_man_B_Madrid	35	1	0	0
ANG_man_C_Segovia	42	1	0	0

ANG_man_C_x	17	4	1	0
ANS_man_C_x	2	0	0	0
ANT_man_B_Madrid	32	0	2	4
ANT_man_x_x	0	0	0	0
ANU_man_x_x	0	0	0	0
APU_man_B_Segovia	0	0	0	0
ARA_woman_A_Madrid	2	1	0	0
ARA_woman_B_Madrid	13	1	0	1
ARA_woman_C_x	4	0	0	2
ARI_man_x_Argentina	13	0	0	1
ART_man_x_x	0	0	0	0
ATR_woman_x_x	0	0	0	0
AVE_man_x_x	10	2	1	0
BAJ_man_x_x	0	0	0	0
BAR_woman_A_Madrid	3	3	0	0
BEA_woman_A_Navarra	0	0	0	1

BEA_woman_B_Navarra	34	1	0	5
BEC_woman_A_Madrid	8	2	0	0
BEL_man_B_Madrid	9	5	2	0
BEL_woman_B_Madrid	8	0	0	1
BEL_woman_x_x	0	0	0	0
BIE_man_x_x	0	0	0	0
BLA_man_x_x	8	0	0	0
BLA_woman_A_Madrid	3	0	0	1
BRE_man_x_x	2	0	0	0
BUS_man_C_x	1	0	0	0
CAL_man_x_x	0	0	0	0
CAM_man_C_Murcia	1	0	0	0
CAM_woman_B_x	0	0	0	0
CAR_man_A_Segovia	0	0	0	6
CAR_man_x_Alicante	0	0	0	0
CAR_man_x_Madrid	0	0	0	1

CAR_woman_B_Madrid	32	3	0	1
CAR_woman_C_Madrid	44	3	0	3
CAR_woman_C_x	1	0	0	0
CAY_man_x_x	0	0	0	2
CEL_woman_x_x	1	0	0	0
CES_man_B_Segovia	3	1	0	1
CES_man_x_x	3	4	0	0
CHA_man_B_Valladolid	33	1	0	0
CHA_woman_x_x	2	0	0	0
CHE_man_A_Ferrol	0	0	0	0
CHI_man_B_Madrid	0	0	0	0
CHI_woman_x_x	1	0	0	0
CIE_man_C_x	10	8	0	0
CLE_woman_C_Segovia	1	0	0	3
COB_man_x_x	0	0	1	0
CON_woman_x_x	0	0	0	0

CRI_man_B_Madrid	25	3	0	3
CRI_woman_A_Madrid	6	0	0	1
CRI_woman_A_Segovia	5	0	2	0
CRI_woman_B_Madrid	21	6	2	0
CRI_woman_B_Segovia	0	0	0	0
CUR_man_x_x	1	0	0	0
DAM_man_x_x	5	0	0	0
DAN_man_B_Argentina	3	0	0	0
DAN_man_B_Madrid	2	0	0	3
DAN_man_x_x	0	0	0	0
DAS_woman_B_Madrid	12	0	0	8
DAV_man_A_Madrid	44	3	0	2
DAV_man_B_Segovia	1	0	0	4
DAV_man_x_x	0	0	0	0
DEB_man_B_Madrid	8	1	0	0
DEF_man_C_Cataluña	25	1	1	1

DIR_man_x_x	0	0	0	0
DOA_woman_A_Egipto	2	1	0	0
DOC_man_x_x	0	0	0	0
DOL_woman_D_Madrid	14	0	0	7
DOS_man_C_Madrid	39	0	1	2
DOS_man_x_Mexico	4	0	0	0
DRA_man_x_x	2	1	0	0
DUD_man_B_Segovia	0	0	0	0
DUR_man_C_Cataluña	3	0	0	0
EDU_man_C_x	12	1	0	0
ELA_woman_x_x	2	0	0	0
ELE_woman_A_Madrid	18	1	1	2
ELE_woman_B_Madrid	2	2	1	2
ELE_woman_C_x	0	0	0	0
ELI_man_x_x	1	0	0	0
EMB_man_x_x	3	3	0	0

EMI_man_A_Madrid	7	2	0	1
EMI_man_x_x	3	0	0	0
ENF_woman_x_x	0	0	0	0
ENR_man_B_Madrid	29	1	0	2
ENT_man_A_x	4	0	0	0
ESQ_man_x_x	0	0	0	0
EST_man_B_x	0	0	0	0
EST_man_C_x	5	0	0	0
EST_man_x_Andalucña	6	0	0	0
EST_woman_B_Madrid	7	0	0	0
EST_woman_x_x	1	0	0	0
EVA_woman_A_Madrid	5	0	1	1
EVO_man_B_Madrid	5	0	0	0
FEL_man_x_x	0	0	0	0
FER_man_A_Madrid	2	0	0	1
FER_man_x_x	15	0	0	2

FIS_man_C_x	11	1	0	3
FRA_man_x_x	3	0	0	0
FRE_man_x_x	0	0	0	0
FUE_man_B_Madrid	8	2	0	0
FUE_woman_A_Madrid	4	2	0	0
FUE_woman_C_x	0	0	0	0
FUE_woman_x_x	7	1	1	0
GAB_man_x_x	12	0	0	0
GAN_man_C_x	0	0	0	0
GAR_man_x_x	10	1	0	0
GAS_man_C_x	0	0	0	0
GAT_man_x_x	0	0	0	0
GEM_woman_A_Madrid	13	0	0	0
GEM_woman_B_Madrid	22	0	2	0
GLO_woman_B_Madrid	8	2	0	0
GRA_man_x_x	3	0	0	0

GUA_man_B_Barcelona	3	0	0	0
GUI_man_A_Madrid	8	2	0	3
GUI_man_B_Madrid	7	0	2	1
GUI_man_x_x	3	0	0	0
GUS_man_A_Argentina	2	0	0	1
GUT_man_x_x	4	0	0	2
HAB_woman_x_x	4	0	0	2
HEC_man_x_x	0	0	0	0
HEI_man_C_Madrid	1	0	0	0
HEL_woman_A_Madrid	42	0	0	11
HER_man_B_Madrid	40	4	1	4
HER_man_x_x	0	0	0	0
HER_woman_x_x	0	0	0	0
HIS_man_A_Madrid	15	6	0	0
HOM_man_C_Madrid	8	1	0	0
HOM_man_x_x	0	0	0	1

HOY_man_B_Burgos	23	4	1	0
IDO_woman_B_Madrid	11	0	0	0
INM_woman_A_Madrid	19	2	0	1
INM_woman_x_x	0	0	0	1
ISA_woman_C_Madrid	2	0	0	0
ISA_woman_x_x	1	0	0	0
ISM_man_C_x	0	0	0	0
IVA_man_A_Cantabria	0	0	0	0
JAI_man_C_x	1	0	0	0
JAU_man_x_x	1	2	0	0
JAV_man_B_Madrid	4	2	0	0
JAV_man_B_Palencia	22	3	1	0
JAV_man_x_x	2	1	0	0
JES_man_A_Madrid	0	0	0	0
JES_man_B_Segovia	23	1	0	2
JES_man_C_x	4	0	0	0

JOA_man_A_Madrid	73	5	3	5
JOA_man_x_Cataluña	0	0	0	0
JOA_man_x_Extremadura	2	2	0	0
JOS_man_A_Segovia	3	1	0	1
JOS_man_B_Madrid	55	7	1	3
JOS_man_C_Madrid	80	6	4	0
JOS_man_C_Segovia	22	4	1	2
JOS_man_D_Madrid	0	0	0	1
JOV_woman_B_Madrid	7	1	0	0
JOV_woman_x_Mexico	1	0	0	0
JUA_man_C_Salamanca	3	0	0	0
JUA_man_C_x	15	5	0	8
JUA_woman_x_x	0	0	0	0
JUE_man_C_x	3	0	0	3
JUL_man_C_x	32	3	3	0
KAJ_man_x_x	1	0	0	0

LAG_man_C_Madrid	3	0	0	0
LAN_woman_A_Madrid	14	2	1	2
LAR_man_x_x	0	0	0	0
LAU_woman_B_Madrid	9	0	0	2
LAU_woman_x_x	1	0	1	0
LET_woman_B_Madrid	21	0	0	3
LIN_woman_x_x	1	0	0	0
LIS_man_x_x	3	4	0	0
LOC_man_C_Madrid	63	2	1	4
LOC_man_C_x	13	0	1	3
LOD_man_x_x	0	0	0	0
LOL_woman_B_Madrid	3	0	0	1
LOL_woman_C_x	0	0	0	0
LOL_woman_x_x	0	1	0	0
LUC_man_B_Zamora	27	0	0	10
LUC_man_C_x	2	0	0	0

LUC_woman_A_Madrid	17	3	0	7
LUI_man_A_Madrid	16	4	0	2
LUI_man_C_Madrid	8	0	0	5
LUI_man_C_Segovia	5	0	0	0
LUI_man_C_x	9	0	1	3
LUI_woman_B_Burgos	19	1	0	0
LUI_woman_C_Burgos	7	0	0	1
LUI_woman_D_Madrid	0	0	0	1
LUZ_woman_C_Madrid	1	0	0	2
MAC_man_x_Mexico	0	0	0	0
MAD_man_x_Mexico	0	0	0	0
MAD_woman_C_Badajoz	18	0	0	0
MAD_woman_C_Madrid	0	0	0	1
MAD_woman_C_Segovia	9	0	0	3
MAF_woman_A_Madrid	35	1	1	2
MAJ_man_x_Mexico	0	0	0	1

MAL_man_x_x	1	0	0	0
MAM_woman_C_Madrid	36	0	1	0
MAM_woman_C_Segovia	0	0	0	11
MAN_man_A_Madrid	15	0	0	1
MAN_man_C_Madrid	21	2	1	0
MAN_man_x_x	7	0	2	0
MAR_man_B_Madrid	6	1	0	5
MAR_man_C_Madrid	30	5	0	0
MAR_man_C_x	21	0	0	4
MAR_woman_A_Madrid	54	3	0	10
MAR_woman_B_Madrid	3	0	0	4
MAR_woman_B_x	0	0	0	0
MAR_woman_C_Madrid	13	2	2	1
MAR_woman_D_Madrid	36	0	0	9
MAR_woman_x_x	6	2	0	0
MAS_man_x_x	0	0	0	0

MAS_woman_A_Madrid	0	0	0	0
MAU_man_D_Burgos	13	1	0	7
MAY_man_C_x	0	0	0	0
MAY_woman_B_Madrid	21	0	1	2
MAY_woman_C_Badajoz	32	3	0	7
MAY_woman_x_x	0	0	0	0
MER_man_x_x	0	0	0	0
MIG_man_A_Madrid	6	1	0	5
MIG_man_B_Madrid	23	5	0	11
MIG_man_C_Segovia	0	0	0	0
MIG_man_x_x	0	0	0	0
MIL_man_x_x	0	0	0	0
MOD_man_C_x	0	0	0	0
MOD_man_x_x	3	0	0	0
MON_man_C_Madrid	20	6	2	1
MON_woman_A_Madrid	21	2	0	7

MON_woman_B_Madrid	35	2	0	13
MOR_man_A_C_ceres	1	0	0	0
MUJ_woman_B_Madrid	0	0	0	0
MUJ_woman_x_x	1	0	0	0
NAD_man_B_Mallorca	0	0	0	0
NAN_woman_A_Madrid	7	1	0	0
NAR_man_x_x	1	0	0	0
NAT_woman_B_Madrid	8	1	0	1
NEN_woman_A_Madrid	3	0	2	2
NIE_woman_B_Madrid	1	0	0	2
NIE_woman_C_Madrid	11	2	0	0
NIV_woman_C_Madrid	5	0	0	4
NOT_man_x_x	3	1	0	1
NUR_woman_B_Madrid	14	2	1	5
OLG_woman_A_Madrid	1	0	0	0
OLG_woman_B_Madrid	9	4	0	1

OTR_man_x_x	1	0	0	0
OTR_woman_A_Madrid	1	0	0	1
OTR_woman_x_x	7	0	0	0
OSO_man_B_Madrid	2	0	2	0
PAC_man_A_Madrid	9	8	0	1
PAC_man_B_Sevilla	11	0	0	1
PAC_man_x_x	9	0	0	0
PAD_man_C_Madrid	1	0	0	1
PAL_woman_A_Madrid	2	0	0	3
PAL_woman_B_Madrid	3	0	1	4
PAN_man_x_x	9	4	1	0
PAP_man_C_Madrid	2	0	0	0
PAP_man_C_Segovia	0	0	0	0
PAR_man_x_x	0	0	0	0
PAS_man_B_Madrid	3	1	1	0
PAS_woman_x_x	0	0	0	0

PAT_man_B_x	0	0	0	0
PAT_woman_A_Madrid	13	4	1	5
PAT_woman_B_Madrid	14	5	2	7
PAT_woman_B_x	0	0	0	0
PAT_woman_C_x	0	0	0	0
PAZ_woman_A_Madrid	6	1	0	3
PED_man_x_x	5	1	0	0
PEN_man_C_x	1	0	0	0
PEP_man_x_x	3	0	0	0
PEP_woman_C_Segovia	9	0	0	0
PEP_woman_x_x	5	0	0	2
PEQ_man_x_x	0	0	0	0
PES_woman_C_Madrid	3	0	0	2
PFE_man_x_x	0	0	0	0
PIB_man_x_x	1	0	0	0
PIE_woman_D_Madrid	4	1	1	1

PIL_woman_B_Ciudad Real	6	1	1	2
PIL_woman_B_Segovia	11	1	0	0
PIL_woman_C_Madrid	9	0	0	0
PIL_woman_x_x	3	2	0	1
PIM_man_x_x	0	0	0	0
PLA_man_C_x	5	0	0	0
PLA_woman_x_x	2	0	0	0
POL_man_C_País Vasco	10	0	0	0
PRE_man_C_x	22	2	3	7
PRI_man_x_Mexico	1	0	0	0
PRI_woman_B_Madrid	1	0	0	0
PRI_woman_C_Madrid	68	2	0	8
PRI_woman_x_x	0	0	0	0
PRO_man_x_x	2	0	0	0
QUI_man_A_Madrid	10	0	0	0
QUI_man_x_x	3	0	0	0

RAF_man_C_x	16	2	1	1
RAM_man_x_x	2	0	0	0
RAT_man_C_x	1	0	0	0
RAU_man_A_Madrid	24	0	0	10
RAU_man_B_Madrid	4	0	0	0
RED_man_C_x	0	0	0	0
REP_woman_x_x	1	0	0	0
RES_man_x_x	1	0	0	0
RIC_man_B_Madrid	4	0	0	0
RIC_man_B_Segovia	30	0	1	1
ROB_man_B_Segovia	15	0	0	6
ROD_man_C_x	2	1	0	0
ROS_woman_B_Madrid	23	4	3	4
SAL_woman_A_Madrid	2	1	0	0
SAN_woman_B_Madrid	0	0	0	0
SAR_man_x_x	0	0	0	0

SAR_woman_A_Madrid	10	2	0	3
SEC_man_B_Madrid	11	2	0	2
SEC_man_C_x	3	0	1	0
SEC_man_x_x	0	0	0	1
SER_man_B_Huelva	3	0	0	0
SER_man_B_Segovia	79	6	1	2
SER_man_C_Barcelona	1	0	0	0
SER_man_x_Mexico	0	0	1	0
SEV_man_C_Valencia	4	2	0	0
SEV_man_C_Valladolid	1	0	0	0
SEV_woman_A_Sevilla	8	1	1	1
SEV_woman_B_Sevilla	0	0	0	1
SEW_woman_x_x	0	0	0	0
SIG_man_D_Barcelona	30	3	4	0
SIL_woman_B_Madrid	12	4	0	0
SOF_woman_B_Asturias	1	1	0	3

SON_woman_B_Madrid	15	4	2	1
SUS_woman_B_Valladolid	30	5	1	3
TEL_man_x_x	2	0	0	0
TER_woman_B_Madrid	1	0	0	0
TIA_woman_B_Madrid	10	0	1	4
TIA_woman_x_x	0	0	0	0
TIE_man_x_x	0	0	0	0
TIO_man_x_x	0	0	0	0
TOD_woman_B_Madrid	0	0	0	0
TOM_man_x_x	1	0	0	0
TOR_man_C_x	6	2	0	0
TRA_man_x_x	1	0	0	0
TRE_man_x_Mexico	0	0	0	0
TRI_woman_A_Madrid	4	3	0	4
UEL_man_B_Madrid	6	0	1	1
UNO_man_C_Madrid	6	0	0	0

USE_man_B_Madrid	2	0	0	0
VAM_man_B_Madrid	0	0	0	0
VAY_man_x_Mexico	0	1	0	0
VER_woman_A_Madrid	11	0	0	1
VIC_man_A_x	2	0	0	0
VIC_man_C_x	3	1	0	0
VIC_woman_C_Madrid	14	2	0	2
VID_man_C_Segovia	0	0	0	10
VIS_woman_C_Cuenca	7	0	0	12
VIT_woman_B_Madrid	5	0	0	0
VRI_man_x_x	2	4	0	0
WAL_man_x_x	1	0	1	0
WOM_woman_x_x	0	0	0	0
XYZ_man_x_x	0	0	0	0
YOL_woman_A_Madrid	3	0	0	0
YOL_woman_B_Segovia	0	0	0	0

YUS_man_B_Segovia	0	0	0	1
ZAS_man_C_Madrid	3	0	0	1
Total	3103	309	91	448
Mean	8.19	0.81	0.24	1.18

Table 106: Japanese necessity, possibility, epistemic, deontic and ambiguous markers absolute frequencies

Speaker	**NEC**	**POSS**	**EPIS**	**DEON**	**AMBG**	**Words**
AKI_woman_B_Tokyo	25	43	48	20	0	5304
ANE_man_A_Hokkaido	1	4	3	2	0	293
AYA_woman_D_Tokyo	0	0	0	0	0	839
AYK_woman_A_Tokyo	10	5	6	9	0	1408
CHI_woman_C_Tokyo	12	3	8	7	0	1217
CHO_man_B_Tokyo	10	3	1	12	0	2120
EMI_woman_D_Tokyo	12	7	8	11	0	1609
EMK_woman_A_Tokyo	4	1	2	3	0	1323
HAR_woman_C_Tokyo	7	9	8	8	0	3252
HID_woman_D_Tokyo	19	10	3	26	0	2674

HIR_man_A_Tokyo	39	24	21	42	0	7546
HOS_man_C_Tokyo	16	16	4	28	0	2038
IMU_woman_A_Tokyo	14	15	18	11	0	3150
INT11_woman_D_Tokyo	2	0	2	0	0	377
INT12_woman_C_Tokyo	0	0	0	0	0	10
INT13_man_B_Shizuoka	3	0	3	0	0	420
INT14_woman_C_Tokyo	0	0	0	0	0	5
INT15_woman_A_Tokyo	0	0	0	0	0	2
INT17_man_B_Tokyo	1	0	0	1	0	152
INT18_man_B_Tokyo	0	0	0	0	0	7
INT19_man_B_Shizuoka	1	0	0	1	0	85
INT21_man_B_Shizuoka	0	0	0	0	0	1
INT22_man_B_Shizuoka	0	1	0	1	0	660
INT23_man_B_Shizuoka	0	0	0	0	0	38
KAN_woman_A_Tokyo	9	6	10	5	0	2057
KAS_woman_D_Shizouka	7	2	3	6	0	839

KAY_woman_C_Fukuoka	5	7	8	4	0	1685
KEN_man_B_Shizuoka	9	25	16	18	0	4761
KSA_woman_A_Tokyo	83	39	44	78	0	8124
KUM_woman_A_Kansai	2	3	4	1	0	1152
MAR_man_C_Shizuoka	1	1	0	2	0	618
MAS_man_B_Nara	10	13	11	12	0	4072
MEG_woman_B_Shizuoka	20	24	8	36	0	4016
MIZ_woman_A_Tokyo	15	5	10	10	0	2552
MOE_woman_D_Tokyo	4	2	1	5	0	915
NAR_woman_A_Tokyo	6	9	11	4	0	1071
NOB_man_C_Tokyo	13	4	3	14	0	2902
OKU_woman_A_Chiba	4	7	8	3	0	683
OSM_man_B_Tokyo	10	7	11	6	0	1665
REI_man_B_Tokyo	17	5	9	13	0	2417
SAH_woman_A_Tokyo	21	8	9	20	0	1829
SAK_man_A_Tokyo	4	6	8	2	0	810

SAT_woman_A_Kansai	5	13	11	7	0	1301
SAY_woman_A_Gunma	2	2	1	3	0	1890
SET_woman_A_Tokyo	8	0	1	7	0	1187
SHI_man_D_Tokyo	13	8	8	13	0	4520
SOT_woman_A_Tokyo	3	5	4	4	0	1132
SSA_woman_D_Tokyo	51	38	11	78	0	9046
SUG_woman_A_Tokyo	12	12	13	11	0	2935
TAK_woman_A_Tokyo	16	6	11	11	0	1709
TMA_woman_D_Shizuoka	22	3	7	18	0	9221
TOM_woman_D_Tokyo	1	2	2	1	0	771
TSU_woman_D_Tokyo	5	7	5	7	0	1299
UME_woman_A_Tokyo	2	5	6	1	0	797
YAM_woman_A_Tokyo	22	19	21	20	0	3715
YAN_man_C_Tokyo	18	35	30	23	0	9410
YOS_man_C_Shizuoka	7	2	3	6	0	1288
YUK_woman_A_Saitama	1	1	1	1	0	757

Total	604	472	444	632	0	127676
Mean	10.41	8.14	7.66	10.90	0.00	2201.31

Table 107: Japanese markers' grammatical class absolute frequencies

Speaker	**AUX**	**ADV**	**ADJ**	**MOOD**
AKI_woman_B_Tokyo	22	42	2	2
ANE_man_A_Hokkaido	3	1	0	1
AYA_woman_D_Tokyo	0	0	0	0
AYK_woman_A_Tokyo	9	4	2	0
CHI_woman_C_Tokyo	7	2	5	1
CHO_man_B_Tokyo	12	1	0	0
EMI_woman_D_Tokyo	13	4	0	2
EMK_woman_A_Tokyo	4	1	0	0
HAR_woman_C_Tokyo	12	2	0	2
HID_woman_D_Tokyo	24	3	0	2
HIR_man_A_Tokyo	43	15	3	2
HOS_man_C_Tokyo	24	2	1	5
IMU_woman_A_Tokyo	17	9	1	2

INT11_woman_D_Tokyo	0	0	2	0
INT12_woman_C_Tokyo	0	0	0	0
INT13_man_B_Shizuoka	0	3	0	0
INT14_woman_C_Tokyo	0	0	0	0
INT15_woman_A_Tokyo	0	0	0	0
INT17_man_B_Tokyo	1	0	0	0
INT18_man_B_Tokyo	0	0	0	0
INT19_man_B_Shizuoka	0	0	0	1
INT21_man_B_Shizuoka	0	0	0	0
INT22_man_B_Shizuoka	1	0	0	0
INT23_man_B_Shizuoka	0	0	0	0
KAN_woman_A_Tokyo	6	8	1	0
KAS_woman_D_Shizouka	7	1	0	1
KAY_woman_C_Fukuoka	7	4	1	0
KEN_man_B_Shizuoka	14	13	1	6
KSA_woman_A_Tokyo	77	37	2	6

KUM_woman_A_Kansai	1	4	0	0
MAR_man_C_Shizuoka	1	0	0	1
MAS_man_B_Nara	13	8	0	2
MEG_woman_B_Shizuoka	36	3	0	5
MIZ_woman_A_Tokyo	11	7	2	0
MOE_woman_D_Tokyo	6	0	0	0
NAR_woman_A_Tokyo	8	7	0	0
NOB_man_C_Tokyo	14	2	1	0
OKU_woman_A_Chiba	4	5	1	1
OSM_man_B_Tokyo	8	4	3	2
REI_man_B_Tokyo	16	6	0	0
SAH_woman_A_Tokyo	22	4	3	0
SAK_man_A_Tokyo	3	7	0	0
SAT_woman_A_Kansai	6	11	0	1
SAY_woman_A_Gunma	2	1	0	1
SET_woman_A_Tokyo	7	1	0	0

SHI_man_D_Tokyo	12	6	0	3
SOT_woman_A_Tokyo	5	3	0	0
SSA_woman_D_Tokyo	77	4	0	8
SUG_woman_A_Tokyo	11	10	2	1
TAK_woman_A_Tokyo	11	5	5	1
TMA_woman_D_Shizuoka	18	4	0	3
TOM_woman_D_Tokyo	1	2	0	0
TSU_woman_D_Tokyo	9	2	0	1
UME_woman_A_Tokyo	4	3	0	0
YAM_woman_A_Tokyo	24	14	1	2
YAN_man_C_Tokyo	27	20	4	2
YOS_man_C_Shizuoka	7	2	0	0
YUK_woman_A_Saitama	1	0	0	1
Total	668	297	43	68
Mean	11.52	5.12	0.74	1.17

Appendix B

Script

B.1 Tagset

```
1  ####### TAGS #######
2  tag = {'Modality': 'm', 'Copula': 'cop', 'Adjective': 'adj', 'Candidate
       ': 'candid'}
3  modtype = {'Necessity': 'NEC', 'Possibility': 'POSS'}
4  subtype = {'Epistemic': 'EPIS', 'Deontic': 'DEON', 'Ambiguous': 'AMBG'}
5  clss = {'Aux': 'AUX', 'Adverb': 'Adverb', 'Adjective': 'Adjective', '
       Imperative': 'mood_IMP', 'Subjunctive': 'mood_SUBJ', 'Potential': '
       mood_POT', 'Mai': 'mood_MAI', 'Copula': 'verb_BE'}
6  value = {'0': '0\%', '30': '30\%', '50': '50\%', '70': '70\%', '100': '
       100\%'}
7  negation = {'No': 'no', 'Yes': 'yes'}
8
9  modtag = '<{} modtype="{}" subtype="{}" class="{}" neg="{}" value="{}">
       '
10 wordtag = '<{} class="{}" neg="{}">'
11 candidtag = '<{} modtype="{}" subtype="{}" class="{}" neg="{}" value
       ="{}" id="1">'
12
13 ## MOOD
14 # SUBJUNCTIVE / IMPERATIVE
15 subjunctive = re.compile(r'([^ ]+)_SUBJ')
16 imperative = re.compile(r'([^ ]+)_IMP')
```

```
17
18 ## ADVERBS
19 necepisadv = '<m modtype="NEC" subtype="EPIS" class="Adverb" neg="no"
       value="100\%">'
20
21 # Multiword adverbs
22 multiadvposs = re.compile(r'(a lo mejor|tal vez)', re.IGNORECASE).
       pattern
23 multiadvnec = re.compile(r'(sin duda|sin falta|sin discusión)', re.
       IGNORECASE).pattern
24
25 ## ADJECTIVES
26 necepisadj = 'modtype="NEC" subtype="EPIS" class="Adjective" neg="no"
       value="100\%">'
27
28 ## AUX
29 # POSS
30 possepisaux = modtag.format(tag['Modality'],modtype['Possibility'],
       subtype['Epistemic'],clss['Aux'],negation['No'],value['50'])
31 possdeonaux = modtag.format(tag['Modality'],modtype['Possibility'],
       subtype['Deontic'],clss['Aux'],negation['No'],value['50'])
32 possambgaux = modtag.format(tag['Modality'],modtype['Possibility'],
       subtype['Ambiguous'],clss['Aux'],negation['No'],value['50'])
33 possepiscandid = modtag.format(tag['Candidate'],modtype['Possibility'],
       subtype['Epistemic'],clss['Aux'],negation['No'],value['50'])
34 possdeoncandid = modtag.format(tag['Candidate'],modtype['Possibility'],
       subtype['Deontic'],clss['Aux'],negation['No'],value['50'])
35 possambgcandid = modtag.format(tag['Candidate'],modtype['Possibility'],
       subtype['Ambiguous'],clss['Aux'],negation['No'],value['50'])
36
37 # NEC
38 necepisaux = modtag.format(tag['Modality'],modtype['Necessity'],subtype
       ['Epistemic'],clss['Aux'],negation['No'],value['100'])
39 necdeonaux = modtag.format(tag['Modality'],modtype['Necessity'],subtype
       ['Deontic'],clss['Aux'],negation['No'],value['100'])
40 necambgaux = modtag.format(tag['Modality'],modtype['Necessity'],subtype
       ['Ambiguous'],clss['Aux'],negation['No'],value['100'])
41 necepiscandid = modtag.format(tag['Candidate'],modtype['Necessity'],
       subtype['Epistemic'],clss['Aux'],negation['No'],value['100'])
```

```
42 necdeoncandid = modtag.format(tag['Candidate'],modtype['Necessity'],
       subtype['Deontic'],clss['Aux'],negation['No'],value['100'])
43 necambgcandid = modtag.format(tag['Candidate'],modtype['Necessity'],
       subtype['Ambiguous'],clss['Aux'],negation['No'],value['100'])
```

B.2 Dictionaries

B.2.1 Spanish

```
1
2  ####### AUXILIARIES, MAIN VERBS #######
3  estar = re.compile(r'(( sido)|soy|eres|es|somos|sois|son|sea|seas|sea|
       seamos|seáis|sean|era|eras|era|éramos|erais|eran|fuera|fueras|fuera
       |fuéramos|fuerais|fueran|fui|fuiste|fue|fuimos|fuisteis|fueron|
       sería|serías|sería|seríamos|seríais|serían|seré|serás|será|seremos|
       seréis|serán|sé|sea|sed|sean)', re.IGNORECASE).pattern
4  mainv = re.compile(r'((\b([^ ]+)?(ar|er|ir|ár|ér|ír)(lo|la|los|las|le|
       les)?(se|me|te|nos)?(lo|la|los|las|le|les)?)|(<cop class="verb_BE"
       neg="no">ser([^ ]+)?</cop>))').pattern
5  clitic = re.compile(r'((lo|la|los|las|le|les)|(se|me|te|nos)(lo|la|los|
       las|le|les)?)', re.IGNORECASE).pattern
6  que = re.compile('que').pattern
7  de = re.compile('de').pattern
8  a = re.compile('a').pattern
9
10 auxlemma = re.compile(r'\b(PODER|DEBER|TENER|HABER|DEJAR)\b').pattern
11 poder = re.compile(r'(((he|ha|has|hay|hemos|habéis|han|haya|hayas|haya|
       hayamos|hayáis|hayan|había|habías|había|habíamos|habíais|habían|
       hubiera|hubieras|hubiera|hubiéramos|hubierais|hubieran|hube|hubiste
       |hubo|hubimos|hubisteis|hubieron|habría|habrías|habría|habríamos|
       habríais|habrían|habré|habrás|habrá|habremos|habréis|habrán|he|haya
       |hayan) podido)|puedo|puedes|puede|podemos|podéis|pueden|pueda|
       puedas|pueda|podamos|podáis|puedan|podía|podías|podía|podíamos|
       podíais|podían|pudiera|pudieras|pudiera|pudiéramos|pudierais|
       pudieran|pude|pudiste|pudo|pudimos|pudisteis|pudieron|podría|
       podrías|podría|podríamos|podríais|podrían|podré|podrás|podrá|
       podremos|podréis|podrán|puede|pueda|puedan)', re.IGNORECASE).
       pattern
12 ir = re.compile(r'(((he|ha|has|hay|hemos|habéis|han|haya|hayas|haya|
       hayamos|hayáis|hayan|había|habías|había|habíamos|habíais|habían|
       hubiera|hubieras|hubiera|hubiéramos|hubierais|hubieran|hube|hubiste
       |hubo|hubimos|hubisteis|hubieron|habría|habrías|habría|habríamos|
       habríais|habrían|habré|habrás|habrá|habremos|habréis|habrán|he|haya
```

```
    |hayan) ido)|voy|vas|va|vamos|vais|van|vaya|vayas|vaya|vayamos|
    vayáis|vayan|iba|ibas|iba|íbamos|ibais|iban|fuera|fueras|fuera|
    fuéramos|fuerais|fueran|fui|fuiste|fue|fuimos|fuisteis|fueron|iría|
    irías|iría|iríamos|iríais|irían|iré|irás|irá|iremos|iréis|irán|vaya
    |vayan)', re.IGNORECASE).pattern
13  deber = re.compile(r'(((he|ha|has|hay|hemos|habéis|han|haya|hayas|haya|
    hayamos|hayáis|hayan|había|habías|había|habíamos|habíais|habían|
    hubiera|hubieras|hubiera|hubiéramos|hubierais|hubieran|hube|hubiste
    |hubo|hubimos|hubisteis|hubieron|habría|habrías|habría|habríamos|
    habríais|habrían|habré|habrás|habrá|habremos|habréis|habrán|he|haya
    |hayan) debido)|debo|debes|debe|debemos|debéis|deben|deba|debas|
    deba|debamos|debáis|deban|debía|debías|debía|debíamos|debíais|
    debían|debiera|debieras|debiera|debiéramos|debierais|debieran|debí|
    debiste|debió|debimos|debisteis|debieron|debería|deberías|debería|
    deberíamos|deberíais|deberían|deberé|deberás|deberá|deberemos|
    deberéis|deberán|debe|deba|deban)', re.IGNORECASE).pattern
14  tener = re.compile(r'(((he|ha|has|hay|hemos|habéis|han|haya|hayas|haya|
    hayamos|hayáis|hayan|había|habías|había|habíamos|habíais|habían|
    hubiera|hubieras|hubiera|hubiéramos|hubierais|hubieran|hube|hubiste
    |hubo|hubimos|hubisteis|hubieron|habría|habrías|habría|habríamos|
    habríais|habrían|habré|habrás|habrá|habremos|habréis|habrán|he|haya
    |hayan) tenido)|tengo|tienes|tiene|tenemos|tenéis|tienen|tenga|
    tengas|tenga|tengamos|tengáis|tengan|tenía|tenías|tenía|teníamos|
    teníais|tenían|tuviera|tuvieras|tuviera|tuviéramos|tuvierais|
    tuvieran|tuve|tuviste|tuvo|tuvimos|tuvisteis|tuvieron|tendría|
    tendrías|tendría|tendríamos|tendríais|tendrían|tendré|tendrás|
    tendrá|tendremos|tendréis|tendrán)', re.IGNORECASE).pattern
15  haber = re.compile(r'(((he|ha|has|hay|hemos|habéis|han|haya|hayas|haya|
    hayamos|hayáis|hayan|había|habías|había|habíamos|habíais|habían|
    hubiera|hubieras|hubiera|hubiéramos|hubierais|hubieran|hube|hubiste
    |hubo|hubimos|hubisteis|hubieron|habría|habrías|habría|habríamos|
    habríais|habrían|habré|habrás|habrá|habremos|habréis|habrán|he|haya
    |hayan) habido)|he|has|hay|hemos|habéis|han|haya|hayas|haya|hayamos
    |hayáis|hayan|había|habías|había|habíamos|habíais|habían|hubiera|
    hubieras|hubiera|hubiéramos|hubierais|hubieran|hube|hubiste|hubo|
    hubimos|hubisteis|hubieron|habría|habrías|habría|habríamos|habríais
    |habrían|habré|habrás|habrá|habremos|habréis|habrán|he|haya|hayan)'
    , re.IGNORECASE).pattern
16  hacer = re.compile(r'(((he|has|hay|hemos|habéis|han|haya|hayas|haya|
```

```
    hayamos|hayáis|hayan|había|habías|había|habíamos|habíais|habían|
    hubiera|hubieras|hubiera|hubiéramos|hubierais|hubieran|hube|hubiste
    |hubo|hubimos|hubisteis|hubieron|habría|habrías|habría|habríamos|
    habríais|habrían|habré|habrás|habrá|habremos|habréis|habrán|he|haya
    |hayan) hecho)|hago|haces|hace|hacemos|hacéis|hacen|haga|hagas|haga
    |hagamos|hagáis|hagan|hacía|hacías|hacía|hacíamos|hacíais|hacían|
    hiciera|hicieras|hiciera|hiciéramos|hicierais|hicieran|hice|hiciste
    |hizo|hicimos|hicisteis|hicieron|haría|harías|haría|haríamos|
    haríais|harían|haré|harás|hará|haremos|haréis|harán|haz|haga|hagan)
    ', re.IGNORECASE).pattern
17  poder2 = re.compile(r'(((he|ha|has|hay|hemos|habéis|han|haya|hayas|haya
    |hayamos|hayáis|hayan|había|habías|había|habíamos|habíais|habían|
    hubiera|hubieras|hubiera|hubiéramos|hubierais|hubieran|hube|hubiste
    |hubo|hubimos|hubisteis|hubieron|habría|habrías|habría|habríamos|
    habríais|habrían|habré|habrás|habrá|habremos|habréis|habrán|he|haya
    |hayan) podido)|puedo|puedes|puede|podemos|podéis|pueden|pueda|
    puedas|pueda|podamos|podáis|puedan|podía|podías|podía|podíamos|
    podíais|podían|pudiera|pudieras|pudiera|pudiéramos|pudierais|
    pudieran|pude|pudiste|pudo|pudimos|pudisteis|pudieron|podría|
    podrías|podría|podríamos|podríais|podrían|podré|podrás|podrá|
    podremos|podréis|podrán|puede|pueda|puedan)', re.IGNORECASE)
18  ir2 = re.compile(r'(((he|ha|has|hay|hemos|habéis|han|haya|hayas|haya|
    hayamos|hayáis|hayan|había|habías|había|habíamos|habíais|habían|
    hubiera|hubieras|hubiera|hubiéramos|hubierais|hubieran|hube|hubiste
    |hubo|hubimos|hubisteis|hubieron|habría|habrías|habría|habríamos|
    habríais|habrían|habré|habrás|habrá|habremos|habréis|habrán|he|haya
    |hayan) ido)|voy|vas|va|vamos|vais|van|vaya|vayas|vaya|vayamos|
    vayáis|vayan|iba|ibas|iba|íbamos|ibais|iban|fuera|fueras|fuera|
    fuéramos|fuerais|fueran|fui|fuiste|fue|fuimos|fuisteis|fueron|iría|
    irías|iría|iríamos|iríais|irían|iré|irás|irá|iremos|iréis|irán|vaya
    |vayan)', re.IGNORECASE)
19  deber2 = re.compile(r'(((he|ha|has|hay|hemos|habéis|han|haya|hayas|haya
    |hayamos|hayáis|hayan|había|habías|había|habíamos|habíais|habían|
    hubiera|hubieras|hubiera|hubiéramos|hubierais|hubieran|hube|hubiste
    |hubo|hubimos|hubisteis|hubieron|habría|habrías|habría|habríamos|
    habríais|habrían|habré|habrás|habrá|habremos|habréis|habrán|he|haya
    |hayan) debido)|debo|debes|debe|debemos|debéis|deben|deba|debas|
    deba|debamos|debáis|deban|debía|debías|debía|debíamos|debíais|
    debían|debiera|debieras|debiera|debiéramos|debierais|debieran|debí|
```

```
    debiste|debió|debimos|debisteis|debieron|debería|deberías|debería|
    deberíamos|deberíais|deberían|deberé|deberás|deberá|deberemos|
    deberéis|deberán|debe|deba|deban)', re.IGNORECASE)
20  tener2 = re.compile(r'(((he|ha|has|hay|hemos|habéis|han|haya|hayas|haya
    |hayamos|hayáis|hayan|había|habías|había|habíamos|habíais|habían|
    hubiera|hubieras|hubiera|hubiéramos|hubierais|hubieran|hube|hubiste
    |hubo|hubimos|hubisteis|hubieron|habría|habrías|habría|habríamos|
    habríais|habrían|habré|habrás|habrá|habremos|habréis|habrán|he|haya
    |hayan) tenido)|tengo|tienes|tiene|tenemos|tenéis|tienen|tenga|
    tengas|tenga|tengamos|tengáis|tengan|tenía|tenías|tenía|teníamos|
    teníais|tenían|tuviera|tuvieras|tuviera|tuviéramos|tuvierais|
    tuvieran|tuve|tuviste|tuvo|tuvimos|tuvisteis|tuvieron|tendría|
    tendrías|tendría|tendríamos|tendríais|tendrían|tendré|tendrás|
    tendrá|tendremos|tendréis|tendrán)', re.IGNORECASE)
21  haberq2 = re.compile(r'((((he|ha|has|hay|hemos|habéis|han|haya|hayas|
    haya|hayamos|hayáis|hayan|había|habías|había|habíamos|habíais|
    habían|hubiera|hubieras|hubiera|hubiéramos|hubierais|hubieran|hube|
    hubiste|hubo|hubimos|hubisteis|hubieron|habría|habrías|habría|
    habríamos|habríais|habrían|habré|habrás|habrá|habremos|habréis|
    habrán|he|haya|hayan) habido)|he|has|hay|hemos|habéis|han|haya|
    hayas|haya|hayamos|hayáis|hayan|había|habías|había|habíamos|habíais
    |habían|hubiera|hubieras|hubiera|hubiéramos|hubierais|hubieran|hube
    |hubiste|hubo|hubimos|hubisteis|hubieron|habría|habrías|habría|
    habríamos|habríais|habrían|habré|habrás|habrá|habremos|habréis|
    habrán|he|haya|hayan) que)', re.IGNORECASE)
22  haberd2 = re.compile(r'((((he|ha|has|hay|hemos|habéis|han|haya|hayas|
    haya|hayamos|hayáis|hayan|había|habías|había|habíamos|habíais|
    habían|hubiera|hubieras|hubiera|hubiéramos|hubierais|hubieran|hube|
    hubiste|hubo|hubimos|hubisteis|hubieron|habría|habrías|habría|
    habríamos|habríais|habrían|habré|habrás|habrá|habremos|habréis|
    habrán|he|haya|hayan) habido)|he|has|hay|hemos|habéis|han|haya|
    hayas|haya|hayamos|hayáis|hayan|había|habías|había|habíamos|habíais
    |habían|hubiera|hubieras|hubiera|hubiéramos|hubierais|hubieran|hube
    |hubiste|hubo|hubimos|hubisteis|hubieron|habría|habrías|habría|
    habríamos|habríais|habrían|habré|habrás|habrá|habremos|habréis|
    habrán|he|haya|hayan) de)', re.IGNORECASE)
23
24
25  ####### PERIPHRASES #######
```

```
26 # NEC aux, no change
27 deberaux = re.compile(r'(\b(\%s)\b \%s)' \%(deber, mainv), re.
       IGNORECASE)
28 deberdeaux = re.compile(r'(\b(\%s)\b \%s \%s)' \%(deber, de, mainv), re
       .IGNORECASE)
29 # POSS aux, change
30 poderaux = re.compile(r'(\b\%s\b \%s)' \%(poder,mainv), re.IGNORECASE)
31 poderauxid = re.compile(r'(\b\%s\b(?! \%s))' \%(poder,mainv), re.
       IGNORECASE)
32 # Aux 'a', no change
33 iraux = re.compile(r'(\b\%s\b \%s \%s)' \%(ir,a,mainv), re.IGNORECASE)
34 irauxid1 = re.compile(r'(\b(\%s)\b \b\%s\b)' \%(ir,a), re.IGNORECASE)
35 irauxid2 = re.compile(r'(\b(\%s)\b(?! a))' \%(ir), re.IGNORECASE)
36 # Aux 'que', change
37 auxque = re.compile(r'(\b(\%s|\%s)\b \%s \%s)' \%(tener,haber,que,mainv
       ), re.IGNORECASE)
38 auxqueid1 = re.compile(r'(\b(\%s|\%s)\b \b\%s\b)' \%(tener,haber,que),
       re.IGNORECASE)
39 auxqueid2 = re.compile(r'(\b(\%s|\%s)\b(?! que))' \%(tener,haber), re.
       IGNORECASE)
40 auxquemain = re.compile(r'(\b((\%s|\%s) )?\b\b(\%s)\b(?!(</m>|</cop>)))
       ' \%(que,a,mainv), re.IGNORECASE)
41 # Aux 'de'
42 haberde = re.compile(r'(\b\%s\b \%s \%s)' \%(haber,de,mainv), re.
       IGNORECASE)
43 haberdeid1 = re.compile(r'(\b\%s\b \b\%s\b)' \%(haber,de), re.
       IGNORECASE)
44 haberdeid2 = re.compile(r'(\b\%s\b(?! de))' \%(haber), re.IGNORECASE)
45 deberdeid = re.compile(r'(\b\%s\b \b\%s\b)' \%(deber,de), re.IGNORECASE
       )
46 deberid = re.compile(r'(\b\%s\b(?! de))' \%(deber), re.IGNORECASE)
47
48
49 ####### ADVERBS, ADJECTIVES #######
50 SPAadjpossdict = {}
51 with open("Files/ADJ_POSS_SPA.txt") as f:
52     for line in f:
53         (keyadjposs, valadjposs) = line.split()
54         SPAadjpossdict[(keyadjposs)] = valadjposs
```

```
55
56 SPAadjnecdict = {}
57 with open("Files/ADJ_NEC_SPA.txt") as f:
58     for line in f:
59         (keyadjnec, valadjnec) = line.split()
60         SPAadjnecdict[(keyadjnec)] = valadjnec
61
62 SPAadvpossdict = {}
63 with open("Files/ADV_POSS_SPA.txt") as f:
64     for line in f:
65         (keyadvposs, valadvposs) = line.split()
66         SPAadvpossdict[(keyadvposs)] = valadvposs
67
68 SPAadvnecdict = {}
69 with open("Files/ADV_NEC_SPA.txt") as f:
70     for line in f:
71         (keyadvnec, valadvnec) = line.split()
72         SPAadvnecdict[(keyadvnec)] = valadvnec
```

B.2.2 Japanese

```
1  ############## MAIN VERBS ##############
2  conjuctive = re.compile(r'<v type="conjuctive">([^ ]+?)</v>').pattern
3  basic = re.compile(r'<v type="basic">([^ ]+?)</v>').pattern
4  irrealis = re.compile(r'<v type="irrealis">([^ ]+?)</v>').pattern
5  taform = re.compile(r'<v type="ta">([^ ]+?)</v>').pattern
6  teform = re.compile(r'<v type="te">([^ ]+?)</v>').pattern
7  chaform = re.compile(r'<v type="cha">([^ ]+?)</v>').pattern
8  conditional = re.compile(r'<v type="conditional">([^ ]+?)</v>').pattern
9  basiccond = re.compile(r'<v type="basic_conditional">([^ ]+?)</v>').
       pattern
10 surunoun = re.compile(r'<v type="surunoun">([^ ]+?)</v>').pattern
11 nai = re.compile(r'(ない|なかった|ないで|じゃない|なくて|はない|がない|
       はなかった|ではない|でない|ません|ませんでした|ませんだった|じゃあ
       りません)').pattern
12 masu=re.compile(r'(ます|ました)').pattern
13 potential = re.compile(r'<v type="potential">([^ ]+?)</v>').pattern
14 kureru = re.compile(r'<v type="kureru">([^ ]+?)</v>').pattern
15 mai = re.compile(r'<v type="mai">([^ ]+?)</v>').pattern
16
17 ############## AUXILIARIES ##############
18 beki=re.compile(r'(べき|べきではない|べきでない|べきではなかった|べきで
       なかった|べからず)').pattern
19 houga = re.compile(r'(ほうがいい|方がいい|ほうがよかった|ほうが良かった
       |方がよかった|方が良かった|ほうがよくない|方が良くない|ほうが良くな
       い|\
20 方がよくない|ほうがよくなかった|ほうが良くなかった|方がよかくなかった|
       方がよかくなかった)').pattern
21 ii=re.compile(r'(いい|よかった|良かった|よくない|良くない|よくなかった|
       良くなかった)').pattern
22 kamawanai=re.compile(r'(いい|よかった|良かった|よくない|良くない|よくな
       かった|良くなかった)').pattern
23 nakya0 = re.compile(r'(なけりゃ|なきゃ|なくちゃ|なくちゃった)').pattern
24 nakya1 = re.compile(r'(なければ|なけりゃ|なきゃ|なくては|なくちゃ|なく
       ちゃった)').pattern
25 nakya2 = re.compile(r'(なりません|ならない|ならなかった|いけません|いけ
       ない|いけなかった|だめ|行けない|行けなかった|行けません)').pattern
26 zaruenai = re.compile(r'(ざるをえない|ざるを得ない|ざるをえなかった|ざ
```

```
    るを得なかった)').pattern
27  yamunai = re.compile(r'(やむをえない|やむを得ない|止むを得ない|止むをえ
    ない)').pattern
28  wake = re.compile(r'(わけには行かない|わけにはいかない|わけに行かない|
    わけにいかない|わけには行かなかった|わけにはいかない|わけに行かなか
    った|\
29  わけにいかなかった)').pattern
30  shinobinai = re.compile(r'(にしのびな(い|かった)|に忍びな(い|かった)|に
    しのびません|にしのびませんでした|に忍びません|に忍びませんでした)'
    ).pattern
31  kaneru = re.compile(r'(かねる|かねます|かねた|かねました|かねて|かねな
    かった)').pattern
32  kanenai = re.compile(r'(かねない|かねません|かねなくて)').pattern
33  shikanai = re.compile(r'(しかない|しかなかった)').pattern ###/つい
34  hoshii = re.compile(r'(ほしい|ほしかった|欲しい|欲しかった)').pattern
35  hoshikunai = re.compile(r'(ほしくない|ほしくなかった|ほしくはない|ほし
    くはなかった|欲しくない|欲しくなかった|欲しくはない|欲しくはなかっ
    た)').pattern
36  tai = re.compile(r'(たい|たかった)').pattern
37  takunai = re.compile(r'(たくない|たくなかった|たくはない|たくはなかった
    )').pattern
38  tsumori=re.compile(r'(つもり)').pattern
39  hazu=re.compile(r'(はず)').pattern
40  moraitai = re.compile(r'(もらいたい|もらいたかった)').pattern
41  moraitakunai = re.compile(r'(もらいたくない|もらいたくなかった)').
    pattern
42  kamosuru = re.compile(r'(かも知れない|かもしれない|かも知れません|かも
    しれません|かも)').pattern
43  dekiru = re.compile(r'(できる|できた|できて|できます|できました|出来る|
    出来た|出来て|出来ます|出来ました)').pattern
44  dekinai = re.compile(r'(できない|できなかった|できなくて|できません|出
    来ない|出来なかった|出来なくて|出来ません)').pattern
45  kudasai = re.compile(r'(ください|ちょうだい|なさい|くださいませんか|下
    さい|下さいませんか)').pattern
46  kotoga=re.compile(r'(ことが)').pattern
47  surubeki=re.compile(r'(すべき)').pattern
48  chigainai=re.compile(r'(にちがいない|にちがいなかった|にちがいません|に
    違いない|に違いなかった|に違いません)').pattern
49  deshou=re.compile(r'((?!何|ん).)(でしょう|だろう)((?!？|！|けど|かな|し
```

```
    |<w type="ENDPART">(か|ね|ねえ|なあ)</w>).?)').pattern
50  kagiranai=re.compile(r'(とはかぎらない|とはかぎらなかった|とはかぎらなく|とは限らない|とは限らなかった|とは限らなく)').pattern
51  hodonokoto=re.compile(r'(ほどの(こと|事)(は|も)ない|ほどの(こと|事)(は|も)なかった|ほどの(こと|事)(は|も)なく|程の(こと|事)(は|も)ない|程の(こと|事)(は|も)なかった|程のこ(こと|事)(は|も)なく)').pattern
52  oyobanai=re.compile(r'(におよばない|におよばなかった|におよばなく|に及ばない|に及ばなかった|に及ばなく)').pattern
53  youni=re.compile(r'(ように|におよばなかった|におよばなく|に及ばない|に及ばなかった|に及ばなく)').pattern
54  ha=re.compile(r'(は)').pattern
55  mo=re.compile(r'(も)').pattern
56
57  ############### SENTENCE ENGINGS ###############
58  endpart = re.compile(r'(<w type="ENDPART">[^ ]+?</w>)').pattern
59  punct = re.compile(r'(。|、|から|けど|と|って|？|！|@@@)').pattern
60  desu = re.compile(r'(だ|だった|じゃない|です|でした|でわありません)').
        pattern
61  desuadj = re.compile(r'(だ|った|です|す|<v type="ta">した</v>)').
        pattern
62  desuadjneg = re.compile(r'((じゃ|は)?ない|(じゃ|は)?なかった|<v type="
        conjuctive">あり</v>ません|わ<v type="conjuctive">あり</v>ません|わ
        <v type="conjuctive">あり</v>ませんでした|<v type="conjuctive">あり
        </v>ませんでした)').pattern
63  comma=re.compile(r'(、)').pattern
```

B.3 Tagger

B.3.1 Spanish

```
1  ##################### LOADED/IMPORTED ELEMENTS ####################
2  from regexmod_SPA import *
3  from xml.sax.saxutils import escape
4  from lxml import etree as ET
5  import re
6  from subprocess import Popen, PIPE, STDOUT
7  import sys
8  import cgitb; cgitb.enable()
9  from requests import post
10 from bs4 import BeautifulSoup
11 from dictionaries_SPA import *
12 from tags_SPA import *
13 import pandas as pd
14 import operator
15
16 # Counter for measuring copula/predicative adjective distance
17 copcnt=0
18
19 # Counters for markers
20 neccnt=0
21 posscnt=0
22 episcnt=0
23 deoncnt=0
24 ambgcnt=0
25 debercnt=0
26 podercnt=0
27 iracnt=0
28 tenercnt=0
29 haberqcnt=0
30 haberdcnt=0
31 auxcnt=0
32 advcnt=0
33 adjcnt=0
34 impcnt=0
```

```
35 subjcnt=0
36
37 # Grampal API
38 gp_en = 'http://cartago.lllf.uam.es:8090/gp/'
39
40 # Flags for multiple/separated items
41 negflag=False
42 mainverbflag=False
43 poderflag=False
44 deberauxflag=False
45 auxqueflag=False
46 haberdeflag=False
47 irflag=False
48 queflag=False
49 deflag=False
50 aflag=False
51 mainvflag=False
52 beflag=False
53
54 # Variable for sentence delimiters
55 sentencedeli = re.compile(r'(([\.\?\!;]+|\.\.\.) |([^0-9]: ))').pattern
56
57 # Temporal file
58 xmltempfile = open('Temp/xmltempfile.xml','w', encoding='utf-8')
59
60 ####### Negative particle file #######
61 with open ('Files/ADV_NEG_SPA.txt', 'r', encoding='utf-8') as advneg:
62     advneg = advneg.read().splitlines()
63
64 ####### Expression file #######
65 with open ('Files/exprnec.txt', 'r', encoding='utf-8') as exprnec:
66     exprnec = exprnec.read().splitlines()
67 with open ('Files/exprposs.txt', 'r', encoding='utf-8') as exprposs:
68     exprposs = exprposs.read().splitlines()
69
70
71 ########################## INPUT ##########################
72 print('\n')
73 print ('#'.center(63,'#'))
```

```
74 print ('\%s'.center(30,'#') \%(' SPANISH MODALITY TAGGER '))
75 print ('#'.center(63,'#'))
76 print ('\n— Input: Sentences in Spanish')
77 print ('— Output: Annotation, classification, and recount of modality\
       nmarkers. Generates an XML tree.')
78 inputini=input('\n— Enter a sentence in Spanish:\n')
79
80 # Tokenisation of sentences
81 inputini=re.sub(sentencedeli,r'\1\n',inputini)
82 print ('<text>', file=xmltempfile)
83
84 ########################## MODAL TAGGER ##########################
85 pretag=[]
86 countertext = []
87 for line in inputini.split('\n'):
88    line=line.strip()
89    if line:
90
91 ####### Tagging the lines with Grampal #######
92       data = {'frase':line}
93       r = post(gp_en, data=data)
94       result=r.text
95       for splitted in result.split('\n'):
96          # Separation of word, lemma
97          word,lemma,*reminder=splitted.split('/')
98          reminder=''.join(reminder)
99          word=word.strip()
100
101 ####### Tagging one word markers: copula, mood, adverbs and predicative
       adjectives #######
102          # Temp tag for copula
103          if 'SER' in lemma:
104             word=word.replace(word,wordtag.format(tag['Copula'],clss['
                   Copula'],negation['No'])+word+'</cop>')
105             beflag=True
106             # Copcnt resets and starts counting when copula is
                   encountered
107             copcnt+=1
108
```

```
109             # Temp tag for imperative
110             elif '2,imper' in reminder:
111                 word=word.replace(word,modtag.format(tag['Modality'],
                        modtype['Necessity'],subtype['Deontic'],clss['
                        Imperative'],negation['No'],value['100'])+word+'</m>')
112                 impcnt+=1
113
114             # Temporal tag for subjunctive
115             elif '2,pres_subj' in reminder and not lemma in auxlemma:
116                 word=word.replace(word,word+'_SUBJ')
117
118             # Tagging predicative adjectives and adverbs
119             elif word in SPAadjpossdict:
120                 if beflag==True and copcnt>0<6:
121                     # Temporal tag "adj" assigned to adjectives as they may
                            or may not be in predicative position
122                     word=word.replace(word,'<adj modtype="POSS" subtype="
                            EPIS" class="Adjective" neg="no" value="'+
                            SPAadjpossdict.get(word)+'">'+word+'</adj>')
123                 else:
124                     word=word
125                     beflag=False
126             elif word in SPAadjnecdict:
127                 if beflag==True and copcnt>0<6:
128                     word=word.replace(word,'<adj modtype="NEC" subtype="EPIS
                            " class="Adjective" neg="no" value="'+SPAadjnecdict.
                            get(word)+'">'+word+'</adj>')
129                 else:
130                     word=word
131                     beflag=False
132             elif word in SPAadvpossdict:
133                 if word:
134                     word=word.replace(word,'<change /><m modtype="POSS"
                            subtype="EPIS" class="Adverb" neg="no" value="'+
                            SPAadvpossdict.get(word)+'">'+word+'</m>')
135                 else:
136                     word=word
137                     beflag=False
138             elif word in SPAadvnecdict:
```

```
139             if word:
140                 word=word.replace(word,'<change /><m modtype="NEC"
                        subtype="EPIS" class="Adverb" neg="no" value="'+
                        SPAadvnecdict.get(word)+'">'+word+'</m>')
141             else:
142                 word=word
143                 beflag=False
144         else:
145             word=word
146             copcnt=0
147             if beflag==True:
148                 beflag=True
149             else:
150                 beflag=False
151
152 ####### Negation #######
153         # Tagging negative elements
154         if word in advneg:
155             pretag.append('<w neg="yes">'+word+'</w>')
156             negflag=True
157
158         ## Subjunctive
159         elif subjunctive.match(word):
160             if negflag==True:
161                 word = re.sub(subjunctive,r'\1',word)
162                 pretag.append(modtag.format(tag['Modality'],modtype['
                        Necessity'],subtype['Deontic'],clss['Subjunctive'],
                        negation['Yes'],value['0'])+word+'</m>')
163                 subjcnt+=1
164             elif negflag==False:
165                 pretag.append(word)
166
167         # Temporal tag for negative elements followed by a pause
168         elif re.search(r'[,;:]',word):
169             if negflag==True:
170                 negflag=False
171                 pretag.append('<stop>'+word+'</stop>')
172             else:
173                 pretag.append(word)
```

```
174             else:
175                 pretag.append(word)
176
177
178 ####### Tagging multiword markers #######
179 # Processing the text as a string
180 text=' '.join(pretag)
181 del pretag[:]
182 # Separation of each sentence in a new line
183 text=re.sub(sentencedeli,r'\1\n',text)
184 text=re.sub(r' ((<stop>)?([\.\?\!¡¿;:,]|\.\.\.))',r'\1',text)
185 # Processing each sentence
186 for sentence in text.split('\n'):
187     sentence=sentence.strip()
188
189     # Multiword adverbs  sentence=re.sub(multiadvnec,modtag.format(tag['
            Modality'],modtype['Necessity'],subtype['Epistemic'],clss['
            Adverb'],negation['No'],value['100'])+r'\1'+'</m>',sentence)
190     sentence=re.sub(multiadvposs,modtag.format(tag['Modality'],modtype['
            Possibility'],subtype['Epistemic'],clss['Adverb'],negation['No'
            ],value['50'])+r'\1'+'</m>', sentence)
191
192     # Periphrases
193     sentence=re.sub(deberaux,necdeonaux+r'\1</m>',sentence)
194     sentence=re.sub(deberdeaux,necepisaux+r'\1</m>',sentence)
195     sentence=re.sub(iraux,necdeonaux+r'\1</m>',sentence)
196     # Those that may change with negation are tagged with the temporal
            tag "change"
197     sentence=re.sub(poderaux,'<change />'+possambgaux+r'\1</m>',sentence
            )
198     sentence=re.sub(auxque,'<change />'+necdeonaux+r'\1</m>',sentence)
199     sentence=re.sub(haberde,'<change />'+necdeonaux+r'\1</m>',sentence)
200
201     # For possible separated
202     sentence=re.sub(poderauxid,'<change />'+possambgcandid+r'\2</candid>
            ',sentence)
203     sentence=re.sub(irauxid2,necdeoncandid+r'\2</candid>',sentence)
204     sentence=re.sub(irauxid1,necdeoncandid+r'\2 a</candid>',sentence)
205     sentence=re.sub(auxqueid2,'<change />'+necdeoncandid+r'\2</candid>',
```

```
        sentence)
206     sentence=re.sub(auxqueid1,'<change />'+necdeoncandid+r'\2 que</
        candid>',sentence)
207     sentence=re.sub(auxquemain,r'<candref ref="1">\1</candref>',sentence
        )
208     sentence=re.sub(haberdeid1,necdeoncandid+r'\2 de</candid>',sentence)
209     sentence=re.sub(haberdeid2,necdeoncandid+r'\2</candid>',sentence)
210     sentence=re.sub(deberdeid,necepiscandid+r'\2 de</candid>',sentence)
211     sentence=re.sub(deberid,necdeoncandid+r'\2</candid>',sentence)
212
213     text='\n<s>'+sentence+'</s>\n'
214
215     print(text, file=xmltempfile)
216 print ('</text>', file=xmltempfile)
217
218
219 xmltempfile.close()
220
221
222 ########################## XML PARSING ##########################
223 ####### XML generation and deletion of non-modal candidates #######
224 xmltempfile2 = open('Temp/xmltempfile2.xml','w', encoding='utf-8')
225 with open ('Temp/xmltempfile.xml', 'r', encoding='utf-8') as xmltemp:
226     soup = BeautifulSoup(xmltemp, "xml")
227
228 ####### Negation and adjectives #######
229 for sentence in soup.find_all('s'):
230     try:
231         for siblings in sentence.w.next_siblings:
232             # Measurement of words between negation and auxiliary / copula
                  and adjective
233             if siblings.name != 'change' and siblings.name != 'cop' and
                  siblings.name != 'adj' and siblings.name != 'candid' and
                  siblings.name != 'm':
234                 distanceneg = len(siblings.split())
235
236             # Negation and distance of adjectives
237             elif siblings.name == 'cop':
238                 for siblings2 in sentence.cop.next_siblings:
```

```
239                 if siblings2.name != 'cop' and siblings2.name != 'adj':
240                     distancecop = len(siblings2.split())
241                 if distanceneg<3:
242                     if distancecop<3:
243                         if siblings2.name == 'adj':
244                             siblings2.name = 'm'
245                             siblings2['neg'] = 'yes'
246                             if siblings2['modtype'] == 'NEC':
247                                 siblings2['modtype'] = 'POSS'
248                                 siblings2['value'] = '50%'
249                             elif siblings2['modtype'] == 'POSS':
250                                 siblings2['modtype'] = 'NEC'
251                                 siblings2['value'] = '0%'
252
253             # Negation and distance of periphrases and adverbs
254             elif siblings.name == 'change':
255                 if distanceneg<3:
256                     siblings.next_sibling['neg'] = 'yes'
257                     if siblings.next_sibling['modtype'] == 'NEC':
258                         siblings.next_sibling['modtype'] = 'POSS'
259                         siblings.next_sibling['value'] = '50%'
260                     elif siblings.next_sibling['modtype'] == 'POSS':
261                         siblings.next_sibling['modtype'] = 'NEC'
262                         siblings.next_sibling['value'] = '0%'
263             elif siblings.name == 'candid':
264                 siblings.next_sibling['neg'] = 'yes'
265             elif siblings.name == 'm':
266                 if distanceneg<3:
267                     siblings['neg'] = 'yes'
268                     if siblings['modtype'] == 'NEC':
269                         siblings['modtype'] = 'NEC'
270                         siblings['value'] = '0%'
271     except:
272         continue
273
274 ####### Separated periphrases #######
275 # Similar steps as with negation
276 for sentence2 in soup.find_all('s'):
277     candmodtype = sentence2.candid['modtype']
```

```
278     candsubtype = sentence2.candid['subtype']
279     candneg = sentence2.candid['neg']
280     for siblings2 in sentence2.candid.next_siblings:
281         if siblings2.name != 'change':
282             distancesep = len(siblings2)
283 
284         if siblings2.name == 'candref':
285             if distancesep<3:
286                 siblings2['modtype'] = candmodtype
287                 siblings2['subtype'] = candsubtype
288                 siblings2['neg'] = candneg
289                 siblings2['class'] = 'AUX'
290                 siblings2.name = 'm'
291                 sentence2.candid.name = 'm'
292                 sentence2.m['id'] = '1'
293 
294 ####### Elliptic periphrases #######
295 # The remaining candidates for separated/elliptic auxiliaries are
        marked as "elli"
296 for sentence in soup.find_all('s'):
297     if sentence.candid:
298         for siblings in sentence.candid.next_siblings:
299             if siblings == '.':
300                 sentence.candid['elli'] = 'yes'
301                 del sentence.candid['id']
302                 sentence.candid.name = 'm'
303 
304 # Stripping unnecessary tags, counting modality
305 for sentence in soup.find_all('s'):
306     invalig_tags = ['change', 'candid', 'changeid', 'candref', 'stop', '
        adj', 'cop']
307     for tag in invalig_tags:
308         for match in soup.findAll(tag):
309             match.replaceWithChildren()
310     for mod in sentence.find_all('m'):
311         clss=mod.get('class')
312         mdtype=mod.get('modtype')
313         sbtype=mod.get('subtype')
314         sep = mod.get('ref')
```

```
315         if re.findall(deber2, mod.text):
316             debercnt+=1
317         elif re.findall(poder2, mod.text):
318             podercnt+=1
319         elif re.findall(ir2, mod.text):
320             iracnt+=1
321         elif re.findall(tener2, mod.text):
322             tenercnt+=1
323         elif re.findall(haberq2, mod.text):
324             haberqcnt+=1
325         elif re.findall(haberd2, mod.text):
326             haberdcnt+=1
327         if clss == 'Adverb':
328             advcnt+=1
329         elif clss == 'Adjective':
330             adjcnt+=1
331         elif clss == 'mood_IMP':
332             impcnt+=1
333         elif clss == 'mood_SUBJ':
334             subjcnt+=1
335         # Avoid counting repeated values
336         if not sep:
337             if mdtype == 'NEC':
338                 neccnt+=1
339             elif mdtype == 'POSS':
340                 posscnt+=1
341             if sbtype == 'EPIS':
342                 episcnt+=1
343             elif sbtype == 'DEON':
344                 deoncnt+=1
345             elif sbtype == 'AMBG':
346                 ambgcnt+=1
347             if clss == 'AUX':
348                 auxcnt+=1
349 # Temporal tree is written
350 print(soup, file=xmltempfile2)
351 xmltempfile2.close()
352
353 ########################## XML CLEANING ##########################
```

```
354 # lxml checks XML syntax
355 tree = ET.parse('Temp/xmltempfile2.xml')
356 root = tree.getroot()
357
358 tree.write('def.xml', encoding='utf-8')
359 print(ET.tostring(root, pretty_print=True, encoding='utf-8').decode("
        utf-8", errors="strict"))
360
361 print('\n')
362 print ('\%s'.center(50,'-') \%(' MODALITY COUNT '))
363 print ('Necessity Modality:\t'+str(neccnt))
364 print ('Possibility Modality:\t'+str(posscnt))
365 print ('-'*50)
366 print ('Epistemic Markers:\t'+str(episcnt))
367 print ('Deontic Markers:\t'+str(deoncnt))
368 print ('Ambiguous Markers:\t'+str(ambgcnt))
369 print ('-'*50)
370 print ('Auxiliaries:\t'+str(auxcnt))
371 print ('Adverbs:\t'+str(advcnt))
372 print ('Adjectives:\t'+str(adjcnt))
373 print ('Imperatives:\t'+str(impcnt))
374 print ('Negative subjunctives:\t'+str(subjcnt))
375 print ('PODER + V:\t'+str(podercnt))
376 print ('IR A + V:\t'+str(iracnt))
377 print ('DEBER + V:\t'+str(debercnt))
378 print ('TENER QUE + V:\t'+str(tenercnt))
379 print ('HABER QUE + V:\t'+str(haberqcnt))
380 print ('HABER DE + V:\t'+str(haberdcnt))
381 print ('-'*50)
382 print('TOTAL MARKERS ='+'\t'+str(posscnt+neccnt))
```

B.3.2 Japanese

```
1  ################ LOADED/IMPORTED ELEMENTS ################
2  from xml.sax.saxutils import escape
3  from lxml import etree as ET
4  from textblob import TextBlob
5  import re
6  from subprocess import Popen, PIPE, STDOUT
7  import re, sys, requests, goslate
8  from progressbar import ProgressBar
9  pbar = ProgressBar()
10 from dictionaries_JAP import *
11 from regexmod_JAP import *
12
13 adverbcnt=0
14 adjectivecnt=0
15 imperativecnt=0
16
17 # Variable for sentence delimiters
18 sentencedeli = re.compile(r'(。|？|！|～|…)').pattern
19
20 ####### Output files #######
21 outf = open('Temp/out1.xml', 'w', encoding= 'utf-8')
22 #xmldeffile = open('output.xml','w', encoding= 'utf-8')
23 xmltempfile = open('Temp/xmltempfile.xml','w', encoding= 'utf-8')
24 print ('<text>', file=xmltempfile)
25
26 ####### Hiragana file #######
27 with open ('Files/HIRAGANA.txt', 'r', encoding= 'utf-8') as hiragana:
28     hiragana = hiragana.read().splitlines()
29
30 ########################## INPUT ##########################
31 print('\n')
32 print ('#'.center(63,'#'))
33 print ('\%s'.center(30,'#') \%(' JAPANESE SENTENCE MODALITY TAGGER '))
34 print ('#'.center(63,'#'))
35 print ('\n— Input: Sentences in Japanese')
36 print ('— Output: Tagging, classification, and counting of modality\
     nmarkers. Generates an XML tree.')
```

```
37 texto1=input('\n— Enter a sentence:\n')
38 texto1=re.sub(sentencedeli,r'\1\n',texto1)
39
40 ########################## MODAL TAGGER ##########################
41 ####### Tagging the lines with Juman #######
42
43 p = subprocess.Popen(['juman', '-b', '-u'], stdout=PIPE, stdin=PIPE,
       stderr=STDOUT)
44 nonlat = bytes(texto1, 'utf-8')
45 outp = p.communicate(input=nonlat)[0]
46 resultado = outp.decode()
47 print ('————————————————————— TAGGING RESULT
       —————————————————————\n')
48 print(resultado)
49 lines = []
50 countertext = []
51 for line in pbar(resultado.split('\n')):
52     line=line.strip()
53     try:
54         word,reading,lemma,tag,info1,surun,info2,info3,info4,conj,info5,
              info6,*remainder=line.split()
55         remainder= ''.join(remainder)
56
57 ####### Tagging modal adverbs #######
58         # Tagging of final particles for filtering
59         if surun == '終助詞':
60             lines.append('<w type="ENDPART">'+word+'</w>')
61         elif tag == '副詞' or tag == '助詞' or lemma == 'できる' or lemma
              == '出来る' or lemma == '限る' or lemma == '及ぶ':
62             if reading in JAPadvnecdict:
63                 lines.append('<m modtype="NEC" value="'+JAPadvnecdict.get(
                      word)+'" subtype="EPIS" class="Adverb">'+word+'</m>')
64                 adverbcnt+=1
65             elif reading in JAPadvpossdict:
66                 lines.append('<m modtype="POSS" value="'+JAPadvpossdict.get
                      (word)+'" subtype="EPIS" class="Adverb">'+word+'</m>')
67                 adverbcnt+=1
68             else:
69                 lines.append(word)
```

```
70
71  ####### Tagging verbal nouns #######
72          elif surun == 'サ変名詞' and reading != 'むり':
73              lines.append('<v type="surunoun">'+word+'</v>')
74
75  ####### Tagging modal adjectives #######
76          elif tag == '形容詞' or tag == '名詞':
77              if lemma in JAPadjnecdict:
78                  lines.append('<w modtype="NEC" value="'+JAPadvnecdict.get(
                        word)+'" subtype="EPIS" class="Adjective">'+word+'</w>'
                        )
79              elif lemma in JAPadjpossdict:
80                  lines.append('<w modtype="POSS" value="'+JAPadjpossdict.get
                        (word)+'" subtype="EPIS" class="Adjective">'+word+'</w>
                        ')
81              elif lemma in JAPadjimpossdict:
82                  lines.append('<w modtype="NEC" value="'+JAPadjimpossdict.
                        get(word)+'" subtype="EPIS" class="Adjective" neg="yes
                        ">'+word+'</w>')
83              else:
84                  lines.append(word)
85
86  ####### Imperatives #######
87          elif tag == '動詞':
88              if conj == '命令形' and reading != 'ください':
89                  kanji=re.search(r'[\u4e00-\u9faf]',word).group()
90                  if kanji:
91                      lines.append('<m modtype="NEC" subtype="NOEPIS" class="
                            mood_IMP" value="100">'+word+'</m>')
92                      imperativecnt+=1
93                  else:
94                      lines.append(word)
95
96  ####### Possible potentials #######
97              elif '可能動詞' and word != 'しれ' and reading != 'しら' in
                    remainder:
98                  lines.append('<v type="potentialtrue">'+word+'</v>')
99                  potentialtrue=(r'<v type="potentialtrue">[^ ]+?</v>') #
                        Used below for replacing, counting
```

```
100
101 ####### Verb inflections #######
102             elif conj == '基本連用形':
103                 if not word in hiragana:
104                     if lemma != 'なる' and reading != 'いけ':
105                         lines.append('<v type="conjuctive">'+word+'</v>')
106                     else:
107                         lines.append(word)
108                 elif lemma == 'する':
109                     lines.append('<v type="conjuctive">'+word+'</v>')
110                 else:
111                     lines.append(word)
112             elif conj == '未然形' and word != 'なら' and word != 'ませ'
                    and reading != 'いけ':
113                 if not word in hiragana:
114                     if lemma != '得る' and lemma != 'える' and reading != '
                            いか':
115                         lines.append('<v type="imperfective">'+word+'</v>')
116                     else:
117                         lines.append(word)
118                 else:
119                     lines.append(word)
120             elif conj == 'タ系連用テ形':
121                 lines.append('<v type="te">'+word+'</v>')
122             elif conj == '基本形':
123                 if reading != 'やむ':
124                     lines.append('<v type="basic">'+word+'</v>')
125                 else:
126                     lines.append(word)
127             elif conj == 'タ形':
128                 lines.append('<v type="ta">'+word+'</v>')
129             elif conj == 'タ系条件形':
130                 lines.append('<v type="conditional">'+word+'</v>')
131             elif conj == '基本条件形':
132                 lines.append('<v type="basic_conditional">'+word+'</v>')
133             elif conj == 'タ系連用チャ形':
134                 lines.append('<v type="cha">'+word+'</v>')
135             else:
136                 lines.append(word)
```

```
137         elif tag == '接尾辞':
138             if lemma == 'する':
139                 if conj == '基本連用形':
140                     lines.append('<v type="conjuctive">'+word+'</v>')
141                 elif conj == 'タ系連用テ形':
142                     lines.append('<v type="te">'+word+'</v>')
143                 elif conj == '基本形':
144                     lines.append('<v type="basic">'+word+'</v>')
145                 elif conj == 'タ形':
146                     lines.append('<v type="ta">'+word+'</v>')
147                 elif conj == 'タ系条件形':
148                     lines.append('<v type="conditional">'+word+'</v>')
149                 elif conj == '基本条件形':
150                     lines.append('<v type="basic_conditional">'+word+'</v>')
151                 elif conj == 'タ系連用チャ形':
152                     lines.append('<v type="cha">'+word+'</v>')
153                 else:
154                     lines.append(word)
155             elif lemma == 'られる':
156                 lines.append('<v type="potential">'+word+'</v>')
157             else:
158                 lines.append(word)
159         elif lemma == '特殊' and word == '\\':
160             lines.append(' ')
161         elif tag == '助動詞' and lemma == 'まい':
162             lines.append('<v type="mai">'+word+'</v>')
163         else:
164             lines.append(word)
165     except:
166         if line == 'EOS':
167             corpusverb= ''.join(lines)
168             del lines[:]
169             for line in corpusverb.split('\n'):
170                 surunounverb=re.search(r'<v type="surunoun">([^ ]+?)</v><v
                        eng=([^ ]+?) type="([^ ]+?)">([^ ]+?)</v>',line)
171                 kamoverb=re.search(r'かも<v type="(conjuctive|imperfective)
                        ">([^ ]+?)</v>',line)
172                 if surunounverb:
173                     corpusverb=re.sub(r'<v type="surunoun">([^ ]+?)</v><v
```

```
                    eng=([^ ]+?) type="([^ ]+?)">([^ ]+?)</v>',r'<v \1
                    type="\4">\2\5</v>',corpusverb)
174             if kamoverb:
175                 corpusverb=re.sub(r'かも<v type="(conjuctive|
                        imperfective)">([^ ]+?)</v>',r'かも\2',corpusverb)
176             corpusverb="".join(corpusverb.split('\n'))
177             texto=corpusverb
178             countertext.append(texto)
179
180 ############## Tagging of inflections + modals ##############
181 ##### Surunoun #####
182             texto=re.sub(surunoun+surubeki,r'<m modtype="NEC" subtype="
                    DEON" class="Suffix">\1\2</m>',texto)
183
184 ##### Conjuctive #####
185             texto=re.sub(conjuctive+tai,r'<m modtype="NEC" subtype="
                    DEON" class="Suffix">\1\2</m>',texto)
186             texto=re.sub(conjuctive+kaneru,r'<m modtype="NEC" subtype="
                    EPIS" class="Periphrasis">\1\2</m>',texto)
187             texto=re.sub(conjuctive+kanenai,r'<m modtype="POSS" subtype
                    ="EPIS" class="Periphrasis">\1\2</m>',texto)
188             texto=re.sub(conjuctive+nakya1+nakya2,r'<m modtype="NEC"
                    subtype="DEON" class="Periphrasis">\1\2\3</m>',texto)
189             texto=re.sub(conjuctive+nakya0,r'<m modtype="NEC" subtype="
                    DEON" class="Periphrasis">\1\2</m>',texto)
190             texto=re.sub(conjuctive+kudasai,r'<m modtype="NEC" subtype
                    ="DEON" class="Periphrasis">\1\2</m>',texto)
191             texto=re.sub(conjuctive+mai,r'<m modtype="NEC" subtype="
                    DEON" class="mood_MAI">\1\2</m>',texto)
192
193 ##### Basic #####
194             texto=re.sub(basic+beki,r'<m modtype="NEC" subtype="DEON"
                    class="Suffix">\1\2</m>',texto)
195             texto=re.sub(basic+shikanai,r'<m modtype="NEC" subtype="
                    DEON" class="Periphrasis">\1\2</m>',texto)
196             texto=re.sub(basic+wake,r'<m modtype="NEC" subtype="DEON"
                    class="Periphrasis">\1\2</m>',texto)
197             texto=re.sub(basic+tsumori,r'<m modtype="NEC" subtype="DEON
                    " class="Periphrasis">\1\2</m>',texto)
```

```
198             texto=re.sub(basic+kotoga+dekiru,r'<m modtype="POSS"
                    subtype="DEON" class="Periphrasis">\1\2\3</m>',texto)
199             texto=re.sub(basic+kotoga+dekinai,r'<m modtype="NEC"
                    subtype="DEON" class="Periphrasis">\1\2\3</m>',texto)
200             texto=re.sub(basic+kamosuru,r'<m modtype="POSS" subtype="
                    EPIS" class="Verb">\1\2</m>',texto)
201             texto=re.sub(basic+chigainai,r'<m modtype="NEC" subtype="
                    EPIS" class="Periphrasis">\1\2</m>',texto)
202             texto=re.sub(basic+kagiranai,r'<m modtype="POSS" subtype="
                    EPIS" class="Periphrasis">\1\2</m>',texto)
203             texto=re.sub(basic+hodonokoto,r'<m modtype="POSS" subtype="
                    EPIS" class="Periphrasis">\1\2</m>',texto)
204             texto=re.sub(basic+oyobanai,r'<m modtype="POSS" subtype="
                    DEON" class="Periphrasis">\1\2</m>',texto)
205
206 ##### Imperfective #####
207             texto=re.sub(imperfective+nakya1+nakya2,r'<m modtype="NEC"
                    subtype="DEON" class="Periphrasis">\1\2\3</m>',texto)
208             texto=re.sub(imperfective+nakya0,r'<m modtype="NEC" subtype
                    ="DEON" class="Periphrasis">\1\2</m>',texto)
209             texto=re.sub(imperfective+zaruenai,r'<m modtype="NEC"
                    subtype="DEON" class="Periphrasis">\1\2</m>',texto)
210             texto=re.sub(imperfective+potential,r'<m modtype="POSS"
                    subtype="DEON" class="mood_POT">\1\2</m>',texto)
211
212 ##### Ta-form #####
213             texto=re.sub(taform+houga,r'<m modtype="NEC" subtype="DEON"
                     class="Periphrasis">\1\2</m>',texto)
214             texto=re.sub(taform+shikanai,r'<m modtype="NEC" subtype="
                    DEON" class="Periphrasis">\1\2</m>',texto)
215             texto=re.sub(taform+tsumori,r'<m modtype="NEC" subtype="
                    DEON" class="Periphrasis">\1\2</m>',texto)
216             texto=re.sub(taform+kamosuru,r'<m modtype="POSS" subtype="
                    EPIS" class="Verb">\1\2</m>',texto)
217             texto=re.sub(taform+chigainai,r'<m modtype="NEC" subtype="
                    EPIS" class="Periphrasis">\1\2</m>',texto)
218
219 ##### Te-form #####
220             texto=re.sub(teform+mo+ii,r'<m modtype="POSS" subtype="DEON
```

```
                ” class=”Periphrasis”>\1\2\3</m>’,texto)
221             texto=re.sub(teform+ii,r’<m modtype=”POSS” subtype=”DEON”
                    class=”Periphrasis”>\1\2</m>’,texto)
222             texto=re.sub(teform+moraitai,r’<m modtype=”NEC” subtype=”
                    DEON” class=”Periphrasis”>\1\2</m>’,texto)
223             texto=re.sub(teform+ha+nakya1,r’<m modtype=”NEC” subtype=”
                    DEON” class=”Periphrasis”>\1\2\3</m>’,texto)
224             texto=re.sub(teform+ha+nakya2,r’<m modtype=”NEC” subtype=”
                    DEON” class=”Periphrasis”>\1\2\3</m>’,texto)
225             texto=re.sub(teform+kudasai,r’<m modtype=”NEC” subtype=”
                    DEON” class=”Periphrasis”>\1\2</m>’,texto)
226             texto=re.sub(teform+hoshii,r’<m modtype=”NEC” subtype=”DEON
                    ” class=”Verb”>\1\2</m>’,texto)
227
228 ##### Cha-form #####
229             texto=re.sub(chaform+nakya2,r’<m modtype=”NEC” subtype=”
                    DEON” class=”Periphrasis”>\1\2</m>’,texto)
230
231 ##### Conditional form #####
232             texto=re.sub(conditional+ii,r’<m modtype=”NEC” subtype=”
                    DEON” class=”Periphrasis”>\1\2</m>’,texto)
233
234 ##### Basic conditional form #####
235             texto=re.sub(basiccond+ii,r’<m modtype=”NEC” subtype=”DEON”
                     class=”Periphrasis”>\1\2</m>’,texto)
236
237 ##### PRE-Negatives #####
238             texto=re.sub(imperfective+nai+houga,r’<m modtype=”NEC”
                    subtype=”DEON” class=”Periphrasis”>\1\2\3</m>’,texto)
239             texto=re.sub(conjuctive+nai+houga,r’<m modtype=”NEC”
                    subtype=”DEON” class=”Periphrasis”>\1\2\3</m>’,texto)
240             texto=re.sub(imperfective+nakya1+ii,r’<m modtype=”NEC”
                    subtype=”DEON” class=”Periphrasis”>\1\2\3</m>’,texto)
241             texto=re.sub(conjuctive+nakya1+ii,r’<m modtype=”NEC”
                    subtype=”DEON” class=”Periphrasis”>\1\2\3</m>’,texto)
242             texto=re.sub(imperfective+nai+mo+ii,r’<m modtype=”POSS”
                    subtype=”DEON” class=”Periphrasis”>\1\2\3\4</m>’,texto)
243             texto=re.sub(conjuctive+nai+mo+ii,r’<m modtype=”POSS”
                    subtype=”DEON” class=”Periphrasis”>\1\2\3\4</m>’,texto)
```

```
244             texto=re.sub(imperfective+nai+ii,r'<m modtype="POSS"
                    subtype="DEON" class="Periphrasis">\1\2\3</m>',texto)
245             texto=re.sub(conjuctive+nai+ii,r'<m modtype="POSS" subtype
                    ="DEON" class="Periphrasis">\1\2\3\4</m>',texto)
246             texto=re.sub(imperfective+nai+wake,r'<m modtype="NEC"
                    subtype="DEON" class="Periphrasis">\1\2\3</m>',texto)
247             texto=re.sub(conjuctive+nai+wake,r'<m modtype="NEC" subtype
                    ="DEON" class="Periphrasis">\1\2\3</m>',texto)
248             texto=re.sub(imperfective+nai+ha+nakya1,r'<m modtype="NEC"
                    subtype="DEON" class="Periphrasis">\1\2\3\4</m>',texto)
249             texto=re.sub(imperfective+nai+ha+nakya2,r'<m modtype="NEC"
                    subtype="DEON" class="Periphrasis">\1\2\3\4</m>',texto)
250             texto=re.sub(conjuctive+nai+ha+nakya2,r'<m modtype="NEC"
                    subtype="DEON" class="Periphrasis">\1\2\3\4</m>',texto)
251             texto=re.sub(conjuctive+nai+ha+nakya2,r'<m modtype="NEC"
                    subtype="DEON" class="Periphrasis">\1\2\3\4</m>',texto)
252             texto=re.sub(imperfective+nai+hoshii,r'<m modtype="NEC"
                    subtype="DEON" class="Verb">\1\2\3</m>',texto)
253             texto=re.sub(conjuctive+nai+hoshii,r'<m modtype="NEC"
                    subtype="DEON" class="Verb">\1\2\3</m>',texto)
254             texto=re.sub(nai+hoshii+r'(\$|[^<])',r'<m modtype="NEC"
                    subtype="DEON" class="Verb">\1\2</m>',texto)
255             texto=re.sub(imperfective+nai+tsumori,r'<m modtype="NEC"
                    subtype="DEON" class="Periphrasis">\1\2\3</m>',texto)
256             texto=re.sub(conjuctive+nai+tsumori,r'<m modtype="NEC"
                    subtype="DEON" class="Periphrasis">\1\2\3</m>',texto)
257             texto=re.sub(imperfective+nai+moraitai,r'<m modtype="NEC"
                    subtype="DEON" class="Periphrasis">\1\2\3</m>',texto)
258             texto=re.sub(conjuctive+nai+moraitai,r'<m modtype="NEC"
                    subtype="DEON" class="Periphrasis">\1\2\3</m>',texto)
259             texto=re.sub(conjuctive+nai+kagiranai,r'<m modtype="POSS"
                    subtype="EPIS" class="Periphrasis">\1\2\3</m>',texto)
260             texto=re.sub(imperfective+nai+kagiranai,r'<m modtype="POSS"
                     subtype="EPIS" class="Periphrasis">\1\2\3</m>',texto)
261             texto=re.sub(conjuctive+nai+hodonokoto,r'<m modtype="POSS"
                    subtype="EPIS" class="Periphrasis">\1\2\3</m>',texto)
262             texto=re.sub(imperfective+nai+hodonokoto,r'<m modtype="POSS
                    " subtype="EPIS" class="Periphrasis">\1\2\3</m>',texto)
263
```

```
264 ##### Mood #####
265             texto=re.sub(r'<v type="potentialtrue">([^ ]+?)</v>',r'<m
                    modtype="POSS" subtype="DEON" class="mood_POT">\1</m>',
                    texto)
266             potentialtrue=(r'<v type="potentialtrue">[^ ]+?</v>') #
                    Usado más adelante para recuento
267
268 ##### General #####
269             texto=re.sub(r'([^てで])'+kudasai,r'\1<m modtype="NEC"
                    subtype="DEON" class="Verb">\2</m>',texto)
270             texto=re.sub(r'([^ただる])'+kamosuru,r'\1<m modtype="POSS"
                    subtype="EPIS" class="Verb">\2</m>',texto)
271             texto=re.sub(yamunai+r'(\$|[^<])',r'<m modtype="NEC"
                    subtype="" class="Verb">\1</m>\2',texto)
272             texto=re.sub(imposs+endpart,r'<m modtype="POSS" subtype="
                    EPIS" class="Adjective" neg="yes">\1</m>\2',texto)
273             texto=re.sub(imposs+punct,r'<m modtype="POSS" subtype="EPIS
                    " class="Adjective" neg="yes">\1</m>\2',texto)
274             texto=re.sub(imposs+desu,r'<m modtype="POSS" subtype="EPIS"
                     class="Adjective" neg="yes">\1</m>\2',texto)
275             texto=re.sub(nec+endpart,r'<m modtype="NEC" subtype="EPIS"
                    class="Adjective">\1</m>\2',texto)
276             texto=re.sub(nec+punct,r'<m modtype="NEC" subtype="EPIS"
                    class="Adjective">\1</m>\2',texto)
277             texto=re.sub(nec+desu,r'<m modtype="NEC" subtype="EPIS"
                    class="Adjective">\1</m>\2',texto)
278             texto=re.sub(poss+endpart,r'<m modtype="POSS" subtype="EPIS
                    " class="Adjective">\1</m>\2',texto)
279             texto=re.sub(poss+punct,r'<m modtype="POSS" subtype="EPIS"
                    class="Adjective">\1</m>\2',texto)
280             texto=re.sub(poss+desu,r'<m modtype="POSS" subtype="EPIS"
                    class="Adjective">\1</m>\2',texto)
281
282 ########## Tagging of modals in overlapping sentences ##########
283             texto=re.sub('\%'+surubeki,r'\%<m modtype="NEC" subtype="
                   DEON" class="Suffix" eli="yes">\1</m>',texto)
284             texto=re.sub('\%'+tai,r'\%<m modtype="NEC" subtype="DEON"
                    class="Suffix" eli="yes">\1</m>',texto)
285             texto=re.sub('\%'+kaneru,r'\%<m modtype="NEC" subtype="EPIS
```

```
        " class="Periphrasis" eli="yes">\1</m>',texto)
286     texto=re.sub('\%'+kanenai,r'\%<m modtype="POSS" subtype="
        EPIS" class="Periphrasis" eli="yes">\1</m>',texto)
287     texto=re.sub('\%'+beki,r'\%<m modtype="NEC" subtype="DEON"
        class="Suffix" eli="yes">\1</m>',texto)
288     texto=re.sub('\%'+shikanai,r'\%<m modtype="NEC" subtype="
        DEON" class="Periphrasis" neg="yes">\1</m>',texto)
289     texto=re.sub('\%'+wake,r'\%<m modtype="NEC" subtype="DEON"
        class="Periphrasis" eli="yes">\1</m>',texto)
290     texto=re.sub('\%'+tsumori,r'\%<m modtype="NEC" subtype="
        DEON" class="Periphrasis" eli="yes">\1</m>',texto)
291     texto=re.sub('\%'+kamosuru,r'\%<m modtype="POSS" subtype="
        EPIS" class="Verb" eli="yes">\1</m>',texto)
292     texto=re.sub('\%'+kotoga+dekiru,r'\%<m modtype="POSS"
        subtype="DEON" class="Periphrasis" eli="yes">\1\2</m>',
        texto)
293     texto=re.sub('\%'+kotoga+dekinai,r'\%<m modtype="NEC"
        subtype="DEON" class="Periphrasis" eli="yes">\1\2</m>',
        texto)
294     texto=re.sub('\%'+nakya1+nakya2,r'\%<m modtype="NEC"
        subtype="DEON" class="Periphrasis" eli="yes">\1\2</m>',
        texto)
295     texto=re.sub('\%'+nakya0,r'\%<m modtype="NEC" subtype="DEON
        " class="Periphrasis" eli="yes">\1</m>',texto)
296     texto=re.sub('\%'+zaruenai,r'\%<m modtype="NEC" subtype="
        DEON" class="Periphrasis" eli="yes">\1</m>',texto)
297     texto=re.sub('\%'+houga,r'\%<m modtype="NEC" subtype="DEON"
         class="Periphrasis" eli="yes">\1</m>',texto)
298     texto=re.sub('\%'+tsumori,r'\%<m modtype="NEC" subtype="
        DEON" class="Periphrasis" eli="yes">\1</m>',texto)
299     texto=re.sub('\%'+mo+ii,r'\%<m modtype="POSS" subtype="DEON
        " class="Periphrasis" eli="yes">\1\2</m>',texto)
300     texto=re.sub('\%'+moraitai,r'\%<m modtype="NEC" subtype="
        DEON" class="Periphrasis" eli="yes">\1</m>',texto)
301     texto=re.sub('\%'+ha+nakya2,r'\%<m modtype="NEC" subtype="
        DEON" class="Periphrasis" eli="yes">\1\2</m>',texto)
302     texto=re.sub('\%'+nakya2,r'\%<m modtype="NEC" subtype="DEON
        " class="Periphrasis" eli="yes">\1</m>',texto)
303     texto=re.sub('\%'+kudasai,r'\%<m modtype="NEC" subtype="
```

```
                DEON" class="Periphrasis" eli="yes">\1</m>',texto)
304             texto=re.sub('\%'+yamunai,r'\%<m modtype="NEC" subtype=""
                    class="Verb" eli="yes">\1</m>',texto)
305             texto=re.sub('\%'+hoshii,r'\%<m modtype="NEC" subtype="DEON
                    " class="Verb" eli="yes">\1</m>',texto)
306
307             texto=re.sub(texto,'<s>\n'+texto+'\n</s>',texto)
308             print(texto, file=xmltempfile)
309
310     else:
311         ...
312
313 print ('</text>',file=xmltempfile)
314 xmltempfile.close()
315
316 ############## Counters ##############
317 countertext= ''.join(countertext)
318 bekicnt=len(re.compile(r'(\%s|\%s|\%s)(\%s|\%s)' \%(basic,surunoun,'\%'
    ,beki,surubeki)).findall(countertext))
319 hougacnt=len(re.compile(r'((\%s|\%s)\%s)|((\%s|\%s|\%s)\%s\%s)' \%(
    taform,'\%',houga,imperfective,conjuctive,'\%',nai,houga)).findall(
    countertext))
320 taraiicnt=len(re.compile(r'(\%s|\%s)\%s' \%(conditional,'\%',ii)).
    findall(countertext))
321 rebaiicnt=len(re.compile(r'(\%s|\%s|\%s)\%s' \%(basiccond,nakya1,'\%',
    ii)).findall(countertext)) #nakya1 para marcar el negativo
322 moiicnt=len(re.compile(r'(\%s|\%s|((\%s|\%s)\%s))\%s?\%s' \%(teform,'\%
    ',conjuctive,imperfective,nai,mo,ii)).findall(countertext))
323 nakya0cnt=len(re.compile(r'(\%s|\%s|\%s|\%s)\%s' \%(imperfective,
    conjuctive,chaform,'\%',nakya0)).findall(countertext))
324 nakya1cnt=len(re.compile(r'(\%s|\%s|\%s|\%s)\%s\%s' \%(imperfective,
    conjuctive,chaform,'\%',nakya1,nakya2)).findall(countertext))
325 if nakya0cnt == 0:
326     nakyacnt=nakya1cnt
327 elif nakya0cnt > 0:
328     nakyacnt = nakya0cnt
329 tehanakyacnt=len(re.compile(r'(\%s|\%s|((\%s|\%s)\%s))\%s(\%s|\%s)' \%(
    teform,'\%',imperfective,conjuctive,nai,ha,nakya1,nakya2)).findall(
    countertext))
```

```
330 zarucnt=len(re.compile(r'(\%s|\%s)\%s' \%(imperfective,'\%',zaruenai)).
        findall(countertext))
331 yamucnt=len(re.compile(r'\%s?\%s' \%('\%',yamunai)).findall(countertext
        ))
332 wakecnt=len(re.compile(r'(\%s|\%s|((\%s|\%s)\%s))\%s' \%(basic,'\%',
        conjuctive,imperfective,nai,wake)).findall(countertext))
333 kanerucnt=len(re.compile(r'(\%s|\%s)\%s' \%(conjuctive,'\%',kaneru)).
        findall(countertext))
334 kanerunaicnt=len(re.compile(r'(\%s|\%s)\%s' \%(conjuctive,'\%',kanenai)
        ).findall(countertext))
335 shikacnt=len(re.compile(r'(\%s|\%s|\%s)\%s' \%(basic,taform,'\%',
        shikanai)).findall(countertext))
336 hoshicnt=len(re.compile(r'(((\%s|\%s)\%s)|\%s)?\%s' \%(conjuctive,
        imperfective,nai,'\%',hoshii)).findall(countertext))
337 taicnt=len(re.compile(r'(\%s|\%s)\%s' \%(conjuctive,'\%',tai)).findall(
        countertext))
338 tsumoricnt=len(re.compile(r'(\%s|\%s|\%s|((\%s|\%s)\%s))\%s' \%(basic,
        taform,'\%',conjuctive,imperfective,nai,tsumori)).findall(
        countertext))
339 moraicnt=len(re.compile(r'(\%s|\%s|((\%s|\%s)\%s))\%s' \%(teform,'\%',
        conjuctive,imperfective,nai,moraitai)).findall(countertext))
340 dekirucnt=len(re.compile(r'(((\%s\%s)|\%s)\%s)' \%(basic,kotoga,'\%',
        dekiru)).findall(countertext))
341 dekinaicnt=len(re.compile(r'(((\%s\%s)|\%s)\%s)' \%(basic,kotoga,'\%',
        dekinai)).findall(countertext))
342 kudasaicnt=len(re.compile(r'((\%s|\%s|\%s)?\%s)' \%(conjuctive,teform,'
        \%',kudasai)).findall(countertext))
343 kamocnt=len(re.compile(r'((\%s|\%s|\%s)?\%s)' \%(basic,taform,'\%',
        kamosuru)).findall(countertext))
344 shikacnt=len(re.compile(r'((\%s|\%s|\%s)?\%s)' \%(basic,taform,'\%',
        shikanai)).findall(countertext))
345 chigainaicnt=len(re.compile(r'((\%s|\%s|\%s)?\%s)' \%(basic,taform,'\%'
        ,chigainai)).findall(countertext))
346 deshoucnt=len(re.compile(r'(\%s)' \%(deshou)).findall(countertext))
347 rareruncnt=len(re.compile(r'(\%s\%s)|\%s' \%(imperfective,potential,
        potentialtrue)).findall(countertext))
348 kagiranaicnt=len(re.compile(r'(\%s|\%s|((\%s|\%s)\%s))\%s' \%(basic,'\%
        ',conjuctive,imperfective,nai,kagiranai)).findall(countertext))
349 hodonokotocnt=len(re.compile(r'(\%s|\%s|((\%s|\%s)\%s))\%s' \%(basic,'
```

```
    \%',conjuctive,imperfective,nai,hodonokoto)).findall(countertext))
350 oyobanaicnt=len(re.compile(r'(\%s|\%s)\%s' \%(basic,'\%',oyobanai)).
    findall(countertext))
351 maicnt=len(re.compile(r'(\%s\%s)' \%(conjuctive,mai)).findall(
    countertext))
352 adjectivecnt=len(re.compile(r'(\%s|\%s|\%s)(\%s|\%s|\%s)' \%(imposs,nec
    ,poss,endpart,punct,desu)).findall(countertext))
353
354 print('\n')
355 print ('\%s'.center(50,'-') \%(' MODALITY COUNT '))
356 print ('Adverbs: ', adverbcnt)
357 print ('Imperatives: ', imperativecnt)
358 print ('Adjectives: ', adjectivecnt)
359 print ('+たい: ',taicnt)
360 print ('+ほしい: ',hoshicnt)
361 print ('+べき: ',bekicnt)
362 print ('+なきゃならない: ',nakyacnt)
363 print ('+ざるをえない: ',zarucnt)
364 print ('+やむをえない: ',yamucnt)
365 print ('+しかない: ',shikacnt)
366 print ('+たほうが: ',hougacnt)
367 print ('+たらいい: ',taraiicnt)
368 print ('+ればいい: ',rebaiicnt)
369 print ('+てもいい: ',moiicnt)
370 print ('+てはなきゃ: ',tehanakyacnt)
371 print ('+わけにはいかない: ',wakecnt)
372 print ('+もらいたい: ',moraicnt)
373 print ('+つもり: ',tsumoricnt)
374 print ('+できる: ',dekirucnt)
375 print ('+できない: ',dekinaicnt)
376 print ('+られる: ',rarerucnt)
377 print ('+かねる: ',kanerucnt)
378 print ('+かねない: ',kanerunaicnt)
379 print ('+かも: ',kamocnt)
380 print ('+ください: ',kudasaicnt)
381 print ('+にちがいない: ',chigainaicnt)
382 print ('+でしょう: ',deshoucnt)
383 print ('+とは限らない: ',kagiranaicnt)
384 print ('+ほどのことはない: ',hodonokotocnt)
```

```
385 print ('+におよばない: ',oyobanaicnt)
386 print ('+まい: ',maicnt)
387
388 print('\n')
389 print ('TOTAL MARKERS = ',adverbcnt+imperativecnt+adjectivecnt+taicnt+
        hoshicnt+nakyacnt+tehanakyacnt\
390 +zarucnt+yamucnt+shikacnt+hougacnt+wakecnt+moraicnt+kamocnt+kanerucnt+
        kanerunaicnt+dekirucnt\
391 +dekinaicnt+tsumoricnt+bekicnt+moiicnt+taraiicnt+rebaiicnt+kudasaicnt+
        chigainaicnt+deshoucnt+rarerucnt+kagiranaicnt\
392 +hodonokotocnt+oyobanaicnt+maicnt)
393 print('\n')
394 print ('\%s'.center(54,'-') \%(' XML OUTPUT '))
395
396 outf.close()
397 adverbcnt=0
398 adjectivecnt=0
399 imperativecnt=0
400
401 ### XML generation and deletion of non-modal verbs
402 parser = ET.XMLParser(remove_blank_text=True)
403 tree = ET.parse('Temp/xmltempfile.xml',parser)
404 root = tree.getroot()
405 for oracion in root.findall('.//s'):
406    ET.strip_tags(oracion, 'v')
407    ET.strip_tags(oracion, 'w')
408
409 tree.write('outputxml.xml', pretty_print=True, encoding= 'utf-8')
410
411 print(ET.tostring(root, pretty_print=True, encoding= 'utf-8').decode("
        utf-8", errors="strict"))
412 print ('\n-- A copy of the tree has been save in "outputxml.xml" \n')
413 print ('\n'+'-- Select input format: (F)ile or (S)entence')
```